

A FURTHER STUDY OF THE RELIABILITY OF ENGLISH ESSAYS

By P. E. VERNON and G. D. MILLICAN

Institute of Education, University of London

I. *Introduction.* II. *Correlations between Markers.* III. *Consistency of Writers.*
IV. *Correlational Analysis.* V. *Reliability of a Quality Scale for Marking at the*
Pass-Fail Borderline. VI. *Objective Tests of English Ability.* VII. *Correlational*
Analysis of Tests and Essays. VIII. *Summary.*

I. INTRODUCTION

A large number of careful investigations into the reliability of English essay marking have demonstrated the wide variations between markers, not only in their over-all standards and distributions, but also in their orders of merit. Since differences between distributions can fairly readily be allowed for or corrected in any large-scale use of the essay, the crucial statistic would seem to be the correlation between markers. The importance of each marker's re-mark reliability has also been stressed by Wiseman (9). But scarcely any British investigators, except Ballard (1) and Finlayson (3), have paid attention to a third aspect of reliability, namely variability of candidates at different essay topics. This article, then, is concerned first with the average size of correlation between markers to be expected in various essay examinations, and secondly with the consistency of candidates. A third point is the possibility of overcoming the unreliability of essays at the adult student level by using objective English tests.

II. CORRELATIONS BETWEEN MARKERS

Correlations between markers range rather widely in different enquiries. The series of researches carried out by the subcommittee of the International Institute Examinations Enquiry (4, 5) seems to have given the impression that marking is so unreliable that the essay should be used, if at all, only with great caution; and, partly in consequence, it largely disappeared from the 11+ selection examination.¹ Wiseman (9), however, found that a combined mark from 4 markers was sufficiently consistent to justify reinstating it at this stage. The I.I.E.E. Report does not actually quote any correlations between markers; but they print the marks given by 5 examiners to 50 University scholarship candidates, and these yield an average coefficient of 0.511. The results from 100 School Certificate candidates, marked by

¹ The Editor has pointed out to us that the essay was falling into disfavour in some areas much earlier, probably as a result of the criticisms by Ballard and others.

8 examiners, appear to be similar (unfortunately the opportunity to compare marks on 6 essay topics was missed). Following up the methods suggested by Burt in an L.C.C. *Report on Junior County Scholarship Examinations*, Cast (2) analysed the marks allotted by 12 teachers (using four different procedures) for essays written by 40 central school girls. For the marks given by the method of General Impression the mean coefficient was 0.496. In the research described below, involving 15 markers of Training College students' essays, the mean was only 0.415. Morrison and Vernon (7), however, in an investigation similar to Cast's, obtained a mean correlation of 0.715 from 5 teachers' General Impression markings of 40 11+ primary school pupils. Finlayson (3) likewise reports 0.738 for 6 markers of 197 unselected children's essays; and Wiseman's two sets of 4 markers of 11+ selection examination essays average 0.705. Finally, in a study preliminary to Peaker's (8) survey of composition ability, two experienced markers obtained correlations of 0.91 and 0.86 in marking 190 and 68 essays by representative groups of 11-year and 15-year pupils.

Four factors probably account for most of these variations. First, the less skilled or mature the writers, the more their essays are marked for mechanics rather than for the relatively subjective qualities of style and thought-content. In the last-mentioned research, tetrachoric correlations were also obtained at the 80th, 50th and 20th percentile splits; and these averaged 0.88, 0.90 and 0.94. Clearly objectivity is very high in distinguishing the poorest writers from the rest, but less good when a borderline is drawn near the top end (as in grammar school selection).

Secondly, some markers are more experienced than others. When over 3,000 similar scripts were marked by pairs drawn from a group of 69 teachers, the mean coefficient dropped from about 0.89 to 0.72. Thirdly, a long essay may be more objectively assessed than a short one. The time allowance at 11+ is 30 to 45 minutes, and 45 to 60 minutes were allowed in the present research; but the time may extend to 3 hours in some high-level examinations. A fourth factor, which has been unaccountably neglected hitherto, is the heterogeneity or range of ability among the writers. If the variance in a representative population such as 11+ candidates is halved among School Certificate candidates, or central school pupils, a correlation between markers of 0.71 in the former would drop to 0.42 in the latter. In Wiseman's and in Morrison and Vernon's studies heterogeneity was probably slightly more restricted than in Finlayson's and Peaker's.

It is quite possible that an appropriate choice of essay topics affects reliability also. But Table I shows that, simply by summing for different investigations the incidence of the four factors just mentioned, a close correspondence is obtained with the size of the correlation between markers.

TABLE I. FACTORS INFLUENCING CORRELATIONS BETWEEN MARKERS OF ESSAYS

Authors	$\bar{r}$	Incidence of Unfavourable Factors			
		Mature Writers	Less Experienced Markers	Short Essays	Homo-geneity
Vernon & Millican	0.415	X		X	X
Cast	0.496		X	X	X
Hartog	0.511	XX			X
Wiseman, Finlayson, Vernon and Morrison, Peaker ..	0.70 to 0.74		X	X	
Peaker	0.88			X	

III. CONSISTENCY OF WRITERS

Ballard states that a typical student, writing a series of essays on different topics, may vary in merit over a range equivalent to 6 mental years ; but he presents no precise evidence. Finlayson obtained a mean correlation (among 6 markers) of 0.691 between 2 essays written by 11+ children. This suggests slightly greater discrepancies between performances at different topics than between different markings of the same topic. Similarly in Peaker's study, the correlation of 0.81 between three topics written by 11+ pupils is lower than the correlation between markers for a single topic.

In 1951, the lecturers in English from some 36 Colleges comprising the London Institute of Education decided to investigate the marking of written English. Since students may fail to qualify as teachers if their written or spoken English is inadequate, the lecturers were particularly concerned about methods of assessment and standards at the bottom end of the scale. After exploring the kinds of written work commonly set in Colleges, the authors chose a list of seven typical essay topics, given here in abbreviated form.

I. Write a critical review of a good book that you have read recently.

II (a) Explain to a 10-11 year child how to pack and dispatch a framed picture ; and (b) State concisely how *one* of the following work : a pressure cooker (and four alternatives).

III. Describe an interesting and genuine 'first experience.'

IV. Assemble the arguments that could be used for and against one of the following propositions in a discussion or a debate : (i) That a tax should be imposed on bicycles (and three alternatives).

V. Freedom and discipline (two alternative educational topics suggested).

VI. Write a letter to a chance acquaintance whom you met on a railway journey and from whom you borrowed a book which you are returning (two alternative subjects suggested).

VII. Précis of a 1,300 word article on 'Prejudice against Foreigners.'

These were set, either in or out of class, at intervals of about a fortnight, to 8 groups of 28 second-year students in men's and women's non-graduate Colleges ; (after eliminating all incomplete sets, slight deficiencies in the numbers from two Colleges were made up by excesses from two others). Each group of essays was first marked by the College lecturer on the 15-point literal scale (A+ to E-) to which they were all accustomed. No 'scheme' of marking was imposed ; but it was suggested that half-marks be awarded for Content (clearness of thought, originality, vocabulary), quarter-marks for Structure, and a quarter for Mechanical Accuracy (spelling and punctuation).

The 224 essays on each topic were next shuffled by Latin Square design into 7 batches of 32, each batch being made up of 4 from each College group. They were then marked by 7 new markers. It was desired to keep the burden of marking as low as possible for each of the 8 lecturers and the 7 independent markers ; and by this scheme each student was marked twice only on each topic. An analysis of variance of the second marks is given in Table II. Note that the design did not allow the calculation of any interactions. Lines (a) and (c) show highly significant differences between the standards of the 7 markers, and between the over-all marks on all topics for the 8 College groups. Line (b) shows that there are also significant differences

between the average marks awarded to the different topics. According to Line (d), the variance between students is not significant; that is, students within any one College do not perform consistently well or badly on all topics. However if we combine College and Student variances, we get an *F*-ratio which is significant at .01, indicating that markers of several essays can distinguish real differences in merit when the more heterogeneous groups of students from all Colleges are considered.

TABLE II. ANALYSIS OF VARIANCE OF INDEPENDENT MARKINGS

Source of Variance	S.S.	d.f.	M. Sq.	F	Significance
(a) Between markers	11882.08	6	1980.35	16.79	ss
(b) Between topics	3892.07	6	648.68	5.50	ss
(c) Between Colleges	17394.75	7	2484.96	21.07	ss
(d) Between students within Colleges ..	19853.07	216	91.91	.78	ns
(c + d)	37247.81	223	167.03	1.42	s = .01
(e) Error	157100.57	1332	117.94		
(f) Total	210122.54	1567			

IV. CORRELATIONAL ANALYSIS

The second markings were converted to +5 to -5 normal distributions with the same mean and variance for each topic and each marker (though not for each College). The first markings by the lecturers were then scaled against these so as to eliminate their variations in standards. Table III gives the mean correlations (a) for all 224 students, (b) with College differences eliminated.¹ The variations between markers' rank orders in assessing such homogeneous essays as those of College students are serious enough; but variations between topics are obviously greater still, as shown in line 2 of the table. Even these coefficients are too high, since the College markers not only applied the same criteria of merit in assessing all topics, but

TABLE III. AVERAGE INTER-CORRELATIONS

Type of Correlation	No. of Coefficients	Mean Coefficients	
		(a) Total Group	(b) Within Group
Same essay, correlation between two markers	7	.509	.415
Same College markers, correlation between different topics	21	.366	.242
Different topics, College marker with independent marker	30	.251	.134
Independent markers, correlation between different topics	21	.224	.096

For $N = 224$, $r = .13$ is significant at 5 per cent., and $r = .17$ at 1 per cent.

¹ Line 1 is the mean of the product-moment correlations for the 7 essays; the remaining lines were obtained by Kelley's (6) average inter-correlation formula, eqn. 171. The accuracy of this technique was checked in various ways. For example, the mean r in Table IV, col. (d), is 0.522; as calculated by correlation of sums it was 0.529.

were also likely to be influenced throughout by their knowledge of their own students. Thus when we compare different markers marking different essays, the within-group coefficients fall below the borderline of significance. The line 4 coefficients of 0.224 and 0.096 correspond, of course, to lines $(c + d)$ and (d) respectively in the analysis of variance in Table II. A consistent English ability, recognizable by different examiners from different small samples of students' work, can barely be said to exist. This throws very grave doubt on the common practice among Civil Service, University Scholarship and other examining bodies, of trying to assess English ability in general from a single essay marked by a single examiner—albeit the essay may be a 3-hour rather than a $\frac{3}{4}$ -hour one. It is much more legitimate at the 11-year level because of the greater heterogeneity of the candidates and the greater dependence of their marks on mechanical accuracy; though even here, as Finlayson points out, it is as desirable to obtain two or more essays from each candidate as to have two or more markers—perhaps even more so.

TABLE IV. CONSISTENCY OF ESSAY TOPICS

Topic	(a)	(b)	(c) (d)		(e)
	Marker	Reliabilities	Correlations with Combined Marks		General Factor Loadings
I. Book Review	.477	.432	.655	.507	.609
II. Objective Explanation ..	.411	.276	.577	.414	.404
III. First Experience	.508	.451	.683	.547	.525
IV. Debating Arguments .. .	.459	.375	.696	.564	.483
V. Education Question .. .	.628	.472	.722	.586	.629
VI. A Letter	.534	.434	.624	.463	.447
VII. Précis	.548	.462	.711	.575	.517
Mean	.509	.415	.667	.522	.516

However, a combination of seven essays, each marked twice, provides a fairly consistent measure, especially when one of the markers is the same throughout. Its over-all reliability (by the Spearman-Brown formula) is 0.829 (total group) or 0.718 (within group). Table IV shows that some of the topics are more reliably marked than others, also that some—not necessarily the same ones—are more highly consistent with this grand total. Cols. (a) and (b) give mean correlations between markers; (c) is the total-group correlation between the dual mark for the essay and the total of 14 marks, and (d) the correlation with the sum of the other six pairs of marks; (e) is the general factor-saturation obtained when the seven pairs of marks are analysed by Spearman's method. Topic V is the best, and II the weakest, on all counts.

V. RELIABILITY OF A QUALITY SCALE FOR MARKING AT THE PASS-FAIL BORDERLINE

It seemed possible that correlations between markers for all topics except No. II would be high enough for sets of specimen scripts to be picked out which, in the view of a consensus of markers, would help to define the Pass-Fail borderline. Twenty-four specimens of each were selected from the 224, roughly 4 of which had received

average marks of *C*—, *D*+, *D*, *D*—, *E*+ and *E*/*E*— at the two previous markings. These were submitted to 4 fresh College lecturers with a request to mark Pass (*D* or better) or Fail (*D*— or below).¹ The two previous, and the four new, marks were submitted to analysis of variance, and the results for Topic I are given in Table V. It will be seen that variance between markers is not significant, whereas variance between students is highly significant. Thus a reasonably clear borderline of Failure can be established in this set. The corresponding *F*-ratios for other topics are shown in Table VI. In each case Student variance is satisfactory. But Marker variance is also rather high, indicating considerable disagreement as to where the borderline should be drawn. With No. V, for example, the numbers of Fail marks awarded by different markers ranged from 3 to 18 out of 24. The average inter-correlations of the 6 markers range from 0.171 to 0.326, but the reliability of the six markers combined reaches 0.552 to 0.733. Considering that these figures refer to exceptionally homogeneous sets of below-average scripts only, they are no lower than might be anticipated. They justified the further selection of 4 specimens from each set of 24—one passed by all 6, one failed by 2, (i.e., regarded as just above the borderline), one failed by 4 (just below the borderline), and one failed by 5 or 6 (clear failure). These 6 sets of 4 specimens were circulated to all London Institute lecturers in English in order to give guidance on the standards to be expected of students, and to help reduce variations in marking standards between Colleges.

TABLE V. ANALYSIS OF VARIANCE OF PASS-FAIL GRADES FOR TOPIC I

Source of Variance	S.S.	d.f.	M. Sq.	F	P
Between students	10.972	23	.476	2.75	.001
Between markers	1.806	5	.361	2.08	> .05
Error	19.861	115	.173		
Total	32.639	143			

TABLE VI. *F*-RATIOS AND MARKER CORRELATIONS FOR PASS-FAIL GRADES

Essay	<i>F</i> -Ratios		$\bar{r}$	R
	Students	Markers		
I	2.75**	2.08	.204	.606
III	2.97**	4.14**	.203	.605
IV	2.85**	6.21**	.172	.555
V	3.04**	8.03**	.171	.552
VI	4.29**	3.62*	.313	.732
VII	4.39**	2.95*	.326	.745

¹ Six additional specimens which received first markings of *A*+ to *C*+ were also supplied in order to give markers an idea of the total range of merit, but were not re-marked. Each topic was dealt with by a different set of 4 markers.

VI. OBJECTIVE TESTS OF ENGLISH ABILITY

As there is so little consistency between different markers of different topics, it seemed possible that objective tests would supply as valid an indication of students' English ability as any one marker's judgment of a few essays. Naturally it was not envisaged that such tests could take the place of essays in examining students; but it was hoped that they might supplement the more subjective types of assessment, also that they might be useful diagnostically at the beginning of training for revealing students whose weaknesses in English require special attention. The following battery was constructed and applied to the same 224 students.

I. *Vocabulary*. Fifty multiple-choice items, 10 mins. E.g.

IMPROVIDENT	generous	improving	unlikely
	ungrateful	thrifless	shameless

II. *Reading Comprehension*. Nine varied passages followed by 3 to 7 multiple-choice questions, 48 questions in all; 20 mins.

III. *Spelling*. Ten multiple-choice items, and 20 sentences containing one wrong word to be picked out and rewritten; 10 mins. E.g.

BURYING GROUND	cemetary	cemmetry	cemetry
	cemetery	cemetory	cemettery

Can you garrantee that this is a genuine Swiss watch? . . .

IV. *Punctuation*. Punctuation marks to be inserted in 20 sentences, 40 marks in all; 12 mins. E.g.

What as the economists might enquire is there to show for this now

The scoring of this creative-response test was extremely tedious, and occasionally subjective.

V. *English Usage*. Twenty sentences with one word omitted, e.g.

Give this book to the child w you think will enjoy reading it.

Ten additional 2-choice items covering similar grammatical difficulties; also 5 sentences to be rewritten in good English (1 or 2 marks each). Total 40 marks, 15 mins. E.g.

Having hunted everywhere for the pencil, it was finally found on the table.

This latter part proved to have the highest validity. But its scoring was somewhat subjective; and in a revision of the test multiple-choice form is now being adopted throughout.

Variations between College group means were small, and showed little correspondence with variations in essay means. This might suggest that the tests have poor validity if they cannot even predict which of a number of College classes is likely to write the best essays, as judged by 7 independent markers. But a reasonable explanation is that, by the time students have reached their fourth term, some Colleges have done more than others to improve, say, spelling or usage. Thus one College may be superior in the tests without being superior at essay-writing. Yet the tests may still be diagnostic of writing ability among the students within any one College. All test scores and essay marks were therefore converted to the same +5 to -5 scale within each College; and all correlations reported below are within-group ones.

The existence of teaching differences between Colleges received some confirmation from analysis of regressions of test scores on total essay marks. With English Usage, for example, the correlations ranged from +0.66 in one College to -0.13 in another.

With small groups (28 in each College) such variations might be attributable mainly to chance. Only those for Spelling and Usage, and for all 5 tests combined, approximated to the .05 level of significance. An example is shown in Table VII.

TABLE VII. ANALYSIS OF VARIANCE IN REGRESSIONS

Source of Variance	S.S.	d.f.	M. Sq.	F	P
Separate Regressions of English Usage on Essay total	145.178	8			
Single Regression for all Colleges	79.317	1			
Difference	65.861	7	9.409	1.80	.08
Residual	1090.822	208	5.244		
Total Variance within Groups	1236.000	216			

VII. CORRELATIONAL ANALYSIS OF TESTS AND ESSAYS

The seven paired marks for the different essays and the five tests were inter-correlated.¹ The essays alone yielded a single general factor covering 27.4% of variance (cf. Table IV), and no significant or meaningful bipolars. The loadings of each test on this factor, and the beta weights for optimum prediction, are shown in Table VIII. Multiple R is 0.637; and the correlation of sums when the 5 tests are equally weighted is 0.624. A combination of the three best tests, namely 2 × Vocabulary + Spelling + Usage (but omitting Punctuation because of its scoring difficulties), gives the slightly lower coefficient of 0.603. Note that these tests take no longer to apply than one essay—three-quarters of an hour—and that their predictive value is better than that of 5 of the 7 essays. Moreover, the essay loadings are derived from double-markings. The mean loading for a singly-marked essay is only 0.435, instead of 0.516, that is, much the same as the loading of the Punctuation test alone. The tests might also be improved by revisions; they were experimental forms only.

TABLE VIII. MULTIPLE CORRELATION OF OBJECTIVE TESTS WITH GENERAL ESSAY FACTOR

Test	Essay Factor Loading	Beta Weight
Vocabulary	.489	.2442
Comprehension	.373	.0575
Spelling	.398	.2030
Punctuation	.423	.1684
Usage	.397	.1681

¹ There seems little point in reporting the factorization of this matrix in detail. The bipolar factors distinguished the essays from the test scores, and the Vocabulary + Comprehension tests from the three tests of English mechanics.

The test for Reading Comprehension, however, was not promising. It was included because English teachers often criticize vocabulary, usage, and other objective items as being too 'bitty' to cover the more complex English skills. But actually it contributes scarcely anything not measured by the simple Vocabulary test, with which it correlates to the extent of 0.520.

The question remains—is there anything common to a number of essays at this level which cannot be covered by objective tests. One way of answering this is to extract a general factor from the tests. The mean variance of the 7 essay marks on this factor is 15.4 per cent., thus leaving a residual covariance of 12.3 per cent. among the essay marks. But probably much of this is attributable to the first markings which, as suggested above, attained fair consistency because the College lecturers knew the students. An additional set of correlations was therefore calculated between the separate essay markings and the combination of $2 \times \text{Vocabulary} + \text{Spelling} + \text{Usage}$.¹ After holding constant the test battery, the residual mean correlations (corresponding closely to residual factor variance) were 0.211 for the first markings and 0.077 for the second. We may conclude that different essays marked by different examiners show so little consistency that an improved battery of tests might well give almost as good an indication of general English writing ability. When, however, a marker, or preferably several markers, become familiar with a wide range of the student's work, he or they do arrive at an over-all or average mark which cannot be adequately predicted by objective test scores.²

VIII. SUMMARY

1. Inadequate correlations between different markers of the same essays chiefly occur when the candidates are selected and therefore homogeneous in ability. They are also lowered when the writers are mature, the essays short, or the markers relatively inexperienced. A still more serious source of inconsistency in assessing English ability is the varying performance of candidates when writing essays on different topics.

2. An experiment was carried out with 224 students in London Institute of Education 2-year Training Colleges. This showed virtually no consistent English ability when different essay topics were marked by different markers. Nevertheless a combination of two persons' markings of seven essays provided a reasonably reliable criterion of such ability.

3. Specimens of below-average essays on each topic were graded for Pass or Fail by six markers. Though there were often wide variations in markers' standards, a moderately reliable 'quality scale' of specimens just above and below the borderline was obtained, to assist in future marking.

4. A short battery of objective tests, of which the most promising were Vocabulary, Spelling and English Usage, was found to be as diagnostic of general English writing ability as a single essay. It was shown, however, that the combined judgment of two or more markers who study a wide range of the students' work does yield a writing ability factor which can be only partially predicted by the tests.

¹ Tetrachoric within-group r 's were obtained, to save time, and the results are similar to, but not quite identical with, those shown in Table III. Thus $\bar{r}$ for second markings was 0.133 instead of 0.096.

² The authors' sincere thanks are due first to the College lecturers who gave and marked the essays, to the 7 independent markers and the 24 Pass-Fail graders; and secondly to the Institute's Subcommittee in English which consistently supported a highly technical investigation, despite its implied criticism of their own assessment procedures. They are further indebted to Mr. K. Lovell (Borough Road College) for assistance in applying the objective tests; and to Dr. D. M. Edwards Penfold (Institute of Education) for carrying out the analysis of variance of independent markings, as well as for her advice on aspects of the design of the experiment.

REFERENCES

1. Ballard, P. B. (1923). *The New Examiner*. London : University of London Press.
2. Cast, B. M. D. (1939-40). 'The efficiency of different methods of marking English composition.' *Brit. J. Educ. Psychol.*, IX, 257-69, X, 49-60.
3. Finlayson, D. S. (1951). 'The reliability of the marking of essays.' *Brit. J. Educ. Psychol.*, XXI, 126-34.
4. Hartog, P., Rhodes, E. C., and Burt, C. (1936). *The Marks of Examiners*. London : Macmillan.
5. Hartog, P., et al. (1941). *The Marking of English Essays*. London : Macmillan.
6. Kelley, T. L. (1924). *Statistical Method*. New York : Macmillan.
7. Morrison, R. L., and Vernon, P. E. (1941). 'A new method of marking English compositions.' *Brit. J. Educ. Psychol.*, XI, 109-19.
8. Peaker, G. F. (1953). 'A sampling design used by the Ministry of Education.' *J. Roy. Stat. Soc. Ser. A.*, CXVI, 140-65.
9. Wiseman, S. (1949). 'The marking of English composition in grammar school selection.' *Brit. J. Educ. Psychol.*, XIX, 200-209.

TESTS OF SIGNIFICANCE IN A FACTOR ANALYSIS OF ARTIFICIAL DATA

By D. N. LAWLEY and Z. SWANSON

Mathematical Institute, University of Edinburgh

I. *Problem.* II. *Procedure:* (a) *Description of Population;* (b) *Construction of Samples;* (c) *Maximum Likelihood Estimation and Calculation of Chi-squared.* III. *Results:* (a) *Results for the Complete Sample;* (b) *Results for the Sub-samples.*

I. PROBLEM

The object of this paper is to verify experimentally the maximum likelihood method of estimating factor loadings, and the χ^2 test of significance associated with it, by the use of artificially constructed 'test' variates. A previous investigation of this kind was carried out by Henrysson (3). He employed a system of 9 variates containing one common factor and took 12 samples each of size 200. Our investigation is smaller in the sense that there are 7 variates and a total sample of only 400 sets of observations, divided into sub-samples of 50. It is, however, of greater generality in that two common factors have been introduced into the system. We also wished to see how well the χ^2 test worked with relatively small samples.

II. PROCEDURE

(a) *Description of Population.* Suppose that $y_1, y_2, z_1, \dots, z_7$ represent independent normal variates with zero means and unit variances. Let variates $x_1, \dots, x_7$ be defined as follows:—

$$\begin{aligned}x_1 &= 4y_1 + 3y_2 + 2z_1 \\x_2 &= 5y_1 + 2y_2 + 2z_2 \\x_3 &= 4y_1 + 4y_2 + 2z_3 \\x_4 &= 4y_1 + 2z_4 \\x_5 &= 4y_1 - 2y_2 + 2z_5 \\x_6 &= 5y_1 - 3y_2 + 2z_6 \\x_7 &= 5y_1 - 4y_2 + 2z_7\end{aligned}$$

Then $x_1, \dots, x_7$ may be thought of as normally distributed test variates depending upon two common factors (resulting from y_1 and y_2) and upon uncorrelated residuals $2z_1, 2z_2, \dots$, representing the combined effects of specific factors and error.

The true or population covariance matrix of the test variates is given by

$$C = MM' + D,$$

where $M' = \begin{bmatrix} 4 & 5 & 4 & 4 & 4 & 5 & 5 \\ 3 & 2 & 4 & 0 & -2 & -3 & -4 \end{bmatrix}$

and D is a diagonal matrix with each (diagonal) element equal to 4. Thus the covariance matrix C is found to be

29	26	28	16	10	11	8
	33	28	20	16	19	17
		36	16	8	8	4
			20	16	20	20
				24	26	28
					38	37
						45

where we omit, as for other symmetric matrices, the sub-diagonal elements. The population correlation matrix is then

1.0000	.8405	.8666	.6644	.3790	.3314	.2215
	1.0000	.8124	.7785	.5685	.5365	.4411
		1.0000	.5963	.2722	.2163	.0994
			1.0000	.7303	.7255	.6667
				1.0000	.8609	.8520
					1.0000	.8948
						1.0000

It will be convenient to work throughout with standardized variates and with correlation coefficients rather than covariances. If we factorize the above correlation matrix we obtain standardized population loadings of the tests in the two common factors. As is well known these factors may be subjected to any arbitrary rotation, and in order to define them uniquely we choose loadings such that $\sum \frac{l_1 l_2}{1 - l_1^2 - l_2^2} = 0$, where l_1, l_2 denote the loadings of any test in the two factors and Σ denotes summation over tests. This definition is in accordance with that adopted when applying the maximum likelihood method of estimation to sample data. The values of l_1 and l_2 are found to be as follows :

l_1	.7071	.8474	.6245	.8927	.8400	.8395	.7805
l_2	.6017	.4009	.7063	.0550	-.3574	-.4359	-.5495

(b) *Construction of Samples.* In order to sample from the population described above, pages 2-9 of Wold's (5) 'Random Normal Deviates' were used. On each of these eight pages 9 of the 10 columns were used to provide a sub-sample of 50 sets of observations of the variates $y_1, y_2, z_1, \dots, z_7$ and hence of $x_1, \dots, x_7$. The eight sub-samples were also combined to form a complete sample of size 400.

Covariance and correlation matrices for $x_1, \dots, x_7$ were found both for the complete sample and for the sub-samples. In calculating these, corrections for the mean were applied, though this was not strictly necessary since the true mean of each variate is zero. Thus the number of degrees of freedom for the variances and covariances is 399 in the case of the complete sample and 49 for each sub-sample.

(c) *Maximum Likelihood Estimation and Calculation of Chi-squared.* The maximum likelihood method of estimating factor loadings has been described, for example, by Lawley (4) and has been well exemplified by Emmett (2). An approximate method has been given for calculating the value of χ^2 used to test the hypothesis as to the number of common factors present. This method uses the squares of the residual covariances or correlations and appears to be sufficiently accurate for large samples. In this paper, however, where the sub-samples are of moderate size, we have adopted

the less arithmetically convenient but more accurate method involving a ratio of determinants.

Let R denote the observed correlation (or covariance) matrix. Let L denote the matrix whose elements are the estimated loadings and V the diagonal matrix whose elements are the residual (or specific) variances. Let R_0 be the 'fitted' correlation (or covariance) matrix given by $R_0 = LL' + V$. Then provided that the elements of L are the exact maximum likelihood estimates we may calculate χ^2 as

$n' \log_e(|R_0|/|R|)$, where $n' = n - \frac{1}{6}(2p + 5) - \frac{2}{3}k$ (see Bartlett (1)), n is the number of degrees of freedom for R , p is the number of tests and k is the number of common factors. The main arithmetical labour in the use of this formula arises from the calculation of the determinant $|R|$. On the other hand $|R_0|$ may be found quite easily, as it may be shown to be equal to $|I + J| \times |V|$, where $J = L'V^{-1}L$ and is a diagonal $k \times k$ matrix.

Unfortunately if only a few iterations of the fitting process have been performed and the elements of L are consequently not exact maximum likelihood estimates, the above expression may give an entirely wrong value for χ^2 and achieve a far worse result than the approximate method. This was very forcibly brought home to the authors when in one of the sub-samples a negative value of χ^2 was at first obtained! The remedy is to use instead of the above the slightly more complicated expression

$$n' \{ \log_e(|R_0|/|R|) + \text{tr}(RR_0^{-1}) - p \},$$

where tr denotes trace or sum of diagonal elements.

By using the result $R_0^{-1} = V^{-1} - V^{-1}L(I + J)^{-1}L'V^{-1}$, it is easy to show that

$$\text{tr}(RR_0^{-1}) - p = \text{tr}\{(I + J)^{-1}[J^2 - L'V^{-1}(R - V)V^{-1}L]\},$$

a particularly simple form for calculation since both J and $L'V^{-1}(R - V)V^{-1}L$ are very nearly diagonal $k \times k$ matrices.

It appears that by using this latter expression a reasonably accurate value for χ^2 may be obtained after only a few iterations. The number of degrees of freedom for χ^2 is $\frac{1}{2}\{(p - k)^2 - p - k\}$.

III. RESULTS

(a) *Results for the Complete Sample.* For the complete sample of 400 the observed correlation matrix R was

1.0000	.8389	.8736	.6880	.4132	.3800	.2487
	1.0000	.8455	.8032	.5975	.5981	.4782
		1.0000	.6599	.3480	.3384	.1804
			1.0000	.7538	.7502	.6828
				1.0000	.8724	.8554
					1.0000	.8854
						1.0000

The maximum likelihood estimates of the loadings l_1 and l_2 in the two common factors were found (with errors not likely to exceed .0003) to be

l_1	.7460	.8846	.7209	.9077	.8307	.8376	.7551
l_2	.5421	.3315	.6219	.0099	-.3980	-.4355	-.5806

These values may be compared with the population values previously given. It is evident that sampling errors are large even for a sample of 400.

Tests of Significance in a Factor Analysis of Artificial Data

To test the hypothesis, which we know to be correct, that two common factors only are present we calculate χ^2 by the method of the preceding section, putting $n = 399$, $p = 7$, $k = 2$ and $n' = 394.5$. We find $\chi^2 = 8.7$ with 8 degrees of freedom, and the probability P that this value should be exceeded by chance is about 0.37. The test thus correctly accepts the hypothesis.

(b) *Results for the Sub-samples.* Maximum likelihood estimates of the common factor loadings were found for each of the eight sub-samples and are given below. The population values are also given alongside for easy comparison.

FIRST FACTOR LOADINGS

Sub-sample	Variate	1	2	3	4	5	6	7
1		.672	.838	.665	.895	.889	.910	.873
2		.707	.898	.727	.872	.693	.606	.483
3		.746	.877	.695	.875	.846	.848	.800
4		.820	.923	.742	.935	.783	.805	.774
5		.877	.922	.876	.925	.762	.807	.740
6		.735	.921	.793	.939	.857	.839	.723
7		.667	.795	.544	.902	.852	.829	.773
8		.766	.849	.682	.883	.898	.913	.792
Population		.707	.847	.625	.893	.840	.839	.781

SECOND FACTOR LOADINGS

Sub-sample	Variate	1	2	3	4	5	6	7
1		.563	.424	.695	.167	-.315	-.172	-.436
2		.553	.237	.584	-.188	-.557	-.616	-.792
3		.571	.388	.624	-.063	-.361	-.446	-.514
4		.429	.208	.623	-.001	-.439	-.498	-.573
5		.351	.108	.431	-.103	-.494	-.515	-.580
6		.604	.293	.547	.041	-.416	-.465	-.639
7		.665	.494	.744	.114	-.407	-.426	-.556
8		.533	.429	.677	.088	-.255	-.352	-.524
Population		.602	.401	.706	.055	-.357	-.436	-.549

The most interesting thing about the above results is the large variability of estimated loadings obtained from samples of moderate size.

For each sub-sample the value of χ^2 , with 8 degrees of freedom, was calculated (putting $n = 49$ and $n' = 44.5$). These values and the corresponding probabilities P are as follows :

Sub-sample	Value of χ^2	Probability P
1	7.6	.47
2	4.9	.77
3	2.7	.95
4	20.4	<.01
5	4.3	.83
6	15.1	.05+
7	11.8	.16
8	3.0	.93

It will be seen that in all but one case χ^2 did not exceed the 5 per cent. level of significance. For sub-sample 4, however, χ^2 was significant at the 1 per cent. level. In this case examination of the residual matrix $R - R_0$ revealed a large residual correlation between variates 2 and 7 after the effect of the two common factors had been removed, and this was sufficient to account for the size of the value of χ^2 . The large residual correlation was due to the fact that z_2 and z_7 , which for sub-sample 4 corresponded to columns 7 and 8 on page 5 of 'Random Normal Deviates,' had a correlation coefficient of 0.408, an exceptionally high value which would occur by chance only about 3 times in 1,000. This correlation erroneously implies the presence of a third common factor.

Apart from this the values of χ^2 seem to be rather more variable than might have been expected, though that for sub-sample 6, which nearly reaches the 5 per cent. level of significance, may be explained by the warning which accompanies the corresponding page of 'Random Normal Deviates,' namely, page 7. If we pool the values of χ^2 from the eight sub-samples we obtain a total χ^2 of 69.8 which, with 64 degrees of freedom, is not significant, the value of P being 0.29.

It is possible that for samples of size 50 the χ^2 test may not be very accurate. It is more likely, however, that aberrations among the values of χ^2 are due to an unlucky choice of pages from Wold's tables. On the whole it seems fair to say that our results are in reasonable agreement with theory.

By fitting just one common factor and finding the residual χ^2 (with 14 degrees of freedom) it would be possible to test the hypothesis that only one common factor is present. This has not been done, since it is quite clear from a cursory inspection of the results that the hypothesis would be decisively rejected. It is intended to make further use of the experimental data in an investigation in which maximum likelihood estimation will be compared with the centroid or 'simple summation' method.¹

REFERENCES

1. Bartlett, M. S. (1950). 'Tests of significance in factor analysis.' *Brit. J. Psychol., Stat. Sect.*, III, 77.
2. Emmett, W. G. (1949). 'Factor analysis by Lawley's method of maximum likelihood.' *Brit. J. Psychol., Stat. Sect.*, II, 90.
3. Henrysson, S. (1950). 'The significance of factor loadings.' *Brit. J. Psychol., Stat. Sect.*, III, 159.
4. Lawley, D. N. (1943). 'The application of the maximum likelihood method to factor analysis.' *Brit. J. Psychol.*, XXXIII, 172.
5. Wold, H. (1948). 'Random Normal Deviates.' *Tracts for Computers*, XXV. Cambridge University Press.

¹ Our thanks are due to Mr. W. G. Emmett and Mr. A. E. G. Pilliner for valuable help with some of the calculations.

THE QUARTIMAX METHOD

AN ANALYTIC APPROACH TO ORTHOGONAL SIMPLE STRUCTURE¹

By JACK O. NEUHAUS and CHARLES WRIGLEY

University of Illinois

I. *Introduction.* II. *Rationale of the Quartimax Method.* III. *The Maximization Function.* IV. *Comparison with Carroll's Analytic Procedure.* V. *Numerical Examples.* VI. *Discussion.* VII. *Summary.*

I. INTRODUCTION

One of the problems basic to factor analysis arises from the fact that an infinite number of matrices F can be found such that FF' will restore the correlations. Hence some further restrictions must be introduced to provide for the selection of one rather than any other. One such restriction requires the selection at each stage of the factor which accounts for a maximum of the variance. This results in the principal axes solution. Although it has attractive mathematical properties, psychologists have usually complained that such solutions (like centroid solutions) are difficult to interpret.

Various alternatives have therefore been suggested. Consider, for example, Burt's method (2) of group factor analysis, Holzinger's bifactor solution (4, ch. 6) and Thurstone's simple structure (6, ch. 14). Two criticisms apply to all. First, each depends to some extent upon the judgment of the investigator, either as to the placement of axes in successive rotations, or as to the logical basis for the grouping of variables. Secondly, each is laborious and ill adapted to mechanized computation.

Different criteria for the factor solution will lead to different mathematical functions. The first problem is accordingly to decide upon the criterion. This must be determined by psychological considerations: it is for the psychologist to specify the type of factoring which he finds most convenient. The nature of the criterion then has to be clarified in such a way that an analytic mathematical treatment becomes possible.

The aim of rotation has sometimes been represented as consisting in the attempt to reduce the complexity of the factorial descriptions of the tests: in other words, that factoring is preferred which concentrates the variance of each test in the smallest

¹ This research was supported in part by the United States Air Force under Contract No. 33-(038)-25726 monitored by Air Force Personnel and Training Research Center. Permission is granted for reproduction, translation, publication, use and disposal in whole and in part by or for the United States Government. We should like to acknowledge our indebtedness to Dr. Lester R. Ford, Jr., for his stimulating discussions of the concept of simple structure, to Kern W. Dickman for his assistance in preparing the electronic computer program for the quartimax method, and to Robert D. Boys for his work in the initial calculation of some numerical examples upon a desk calculator. The editorial comments and criticisms of Professor Sir Cyril Burt and Mr. A. Summerfield have also proved most helpful.

number of factors. The ideal rotational case would then consist in a unifactorial factor pattern (4, p. 95) in which the variance for any test is represented by a single loading. This is unattainable with empirical matrices; we shall most nearly approach it if we seek the greatest degree of inequality in the distributing of that variance among the factors. Thus, the aim will be to find the orthogonal transformation which maximizes the variance of the factor contributions (i.e., the squared factor loadings). In this paper, we present a mathematical procedure, called the 'quartimax method,'¹ designed to do this. Since the total test variance must remain constant, an incidental consequence of the quartimax method will be to increase the number of zero or near-zero loadings as well as the size of the larger loadings. To that extent the method forms an approach to 'orthogonal simple structure.'

II. RATIONALE OF THE QUARTIMAX METHOD

Let $F = \{f_{ij}\}$, ($i = 1, \dots, n$; $j = 1, \dots, m$) denote any factoring of R , where n is the number of variables, and m is the number of factors extracted. Let $G = \{g_{ij}\}$ be the factor loading matrix with maximum variance of squared elements. We want to find the orthogonal rotation matrix T such that $FT = G$.

The variance of the squared elements of F is

$$\sigma^2 = \frac{\sum_i \sum_j f_{ij}^4 - \frac{[\sum_i (\sum_j f_{ij}^2)]^2}{mn}}{mn} = \frac{1}{mn} \sum_i \sum_j f_{ij}^4 - \frac{1}{m^2 n^2} [\sum_i \sum_j g_{ij}^2]^2, \quad (1)$$

since, for any row i , $\sum_j g_{ij}^2 = \sum_j f_{ij}^2$. To obtain G therefore we have to find the orthogonal transformation which maximizes the sum of fourth powers of the elements in the rotated matrix. Hence the name 'quartimax.'

In selecting the fourth power for maximization two separate decisions have been made. We have taken the aim to be the reduction of the complexity of the factorial descriptions, and have chosen the sum of fourth powers as our criterion. Other criteria might have been considered. If the essential aim is taken to be to attain a positive manifold, third powers should be maximized. For those who prefer them, the quartimax principle could be extended to the case of oblique transformations. But until psychologists agree on the purpose of rotation, no final solution is possible. In the meanwhile the quartimax method may be useful (a) in showing the practicability of an analytic solution, (b) in accelerating computation, (c) in examining the rotational problem and the relations between current methods, and (d) in encouraging reconsideration of the objectives.

A second reservation should be made. The computational procedures for considering factors two at a time do not necessarily give the highest possible sum. The order can alter the results. Until a procedure is developed for operating on all factors at once, the method given here must suffice for deriving an approximation.²

¹In the previous publications the method was called the 'quadrimax': we are indebted to Sir Cyril Burt for suggesting the term 'quartimax,' which seems a more accurate name.

²Mr. Arthur Gregory of the University of Illinois has considered the problem of getting a direct solution for the full set of m factors. The direct method leads to a non-linear system of equations of the third degree. It may be practicable to approximate to the solution of these equations with the aid of an electronic computer.

III. THE MAXIMIZATION FUNCTION

Let F be the factor-matrix found by the preliminary analysis ; G the matrix for which the sum of the fourth powers of the elements is largest ; and T be the orthogonal matrix which transforms F to G .

A two-by-two transformation matrix will have the form $\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$, and depends only on one parameter θ . The rotated factor loadings can be written :

$$\begin{cases} f'_{ki} = f_{ki} \cos \theta + f_{kj} \sin \theta \\ f'_{kj} = -f_{ki} \sin \theta + f_{kj} \cos \theta \end{cases} \quad (2)$$

Let $h_{ij}(\theta)$ be the sum of the fourth powers of the loadings of the two rotated factors. Our procedure will be to take factors two at a time, find the rotation maximizing $h_{ij}(\theta)$, and so transform the loadings. A complete cycle of operations will involve $\frac{m(m-1)}{2}$ pairings of factors. This will give the factor loading matrix $FT_1 \dots T_i \dots T_{m(m-1)/2}$. No transformation will be made unless the fourth-power sum increases. Since successive iterations give non-decreasing sums of fourth powers throughout the matrix, and the sum of fourth powers must be equal to or less than n , the procedure must converge. Iteration will continue until the desired level of accuracy is reached.¹

Let us now find the value of θ for which $h_{ij}(\theta)$ is a maximum.

$$\begin{aligned} h_{ij}(\theta) &= \sum_{k=1}^n [(f_{ki} \cos \theta + f_{kj} \sin \theta)^4 + (-f_{ki} \sin \theta + f_{kj} \cos \theta)^4] \\ &= \sin 4\theta \cdot \sum_{k=1}^n [f_{ki} f_{kj} (f_{ki}^2 - f_{kj}^2)] - \frac{1}{2} \sin^2 2\theta \cdot \sum_{k=1}^n [(f_{ki}^2 - f_{kj}^2)^2 \\ &\quad - 4f_{ki}^2 f_{kj}^2] + \sum_{k=1}^n (f_{ki}^4 + f_{kj}^4). \end{aligned} \quad (3)$$

To obtain maxima and minima, we find the first derivative $h'_{ij}(\theta)$ of $h_{ij}(\theta)$, set it equal to zero, and solve. Thus we get

$$\begin{aligned} h'_{ij}(\theta) &= -h'_{ij}(\theta + \frac{\pi}{4}) = 4 \cos 4\theta \cdot \sum_{k=1}^n [f_{ki} f_{kj} (f_{ki}^2 - f_{kj}^2)] \\ &\quad - \sin 4\theta \sum_{k=1}^n [(f_{ki}^2 - f_{kj}^2)^2 - 4f_{ki}^2 f_{kj}^2]. \end{aligned} \quad (4)$$

Setting $h'_{ij}(\theta) = 0$, we obtain

$$\theta = \frac{1}{4} \arctan \frac{4 \sum_{k=1}^n [f_{ki} f_{kj} (f_{ki}^2 - f_{kj}^2)]}{\sum_{k=1}^n [(f_{ki}^2 - f_{kj}^2)^2 - 4f_{ki}^2 f_{kj}^2]}. \quad (5)$$

It should be noted that if $\sum_k [f_{ki} f_{kj} (f_{ki}^2 - f_{kj}^2)] = 0$, then $\theta = 0$, i.e., no rotation is required.

Since $\sin 4\theta$ and $\sin^2 2\theta$ both have the period $\frac{\pi}{2}$, so will $h_{ij}(\theta)$. Hence we need only consider rotations θ where $0^\circ \leq \theta < 90^\circ$. In that interval there will be two values for θ . One will correspond to a maximum and the other, 45° away, to a minimum,

¹ Although this computational procedure appears to ensure convergence of the fourth-power sum, it must be remembered that the sum converges to a limit that need not necessarily be the maximum obtainable by a direct solution ; that is to say, the transformed matrix $FT_1 T_2 \dots T_q$ is only an approximation to G .

The Quartimax Method

since $h_{ij}(\theta) = -h_{ij}(\theta + \pi/4) + 3/2 \sum_{k=1}^n [(f_{ki}^2 + f_{kj}^2)^2] = -h_{ij}(\theta + \pi/4) + \text{constant}$. To find whether a given solution θ gives a maximum, θ may be substituted in the second derivative $h''_{ij}(\theta)$ of $h_{ij}(\theta)$. Let $a = \sum_{k=1}^n [f_{ki}f_{kj}(f_{ki}^2 - f_{kj}^2)]$ and $b = \sum_{k=1}^n [(f_{ki}^2 - f_{kj}^2)^2 - 4f_{ki}^2 f_{kj}^2]$; then the second derivative is

$$h''_{ij}(\theta) = -16a \sin 4\theta - 4b \cos 4\theta. \quad (6)$$

If $h''_{ij}(\theta) < 0$, we have a maximum.

It is not necessary to compute eqn. (6) in order to find which angle gives the maximum. The signs of a and b indicate the range of the angle. The conditions for a maximum, viz., $h'_{ij}(\theta) = 4a \cos 4\theta - b \sin 4\theta = 0$

$$\text{and } \frac{h''_{ij}(\theta)}{4} = -b \cos 4\theta - 4a \sin 4\theta < 0$$

$$\text{give } -\frac{b^2 + 16a^2}{4a} \sin 4\theta < 0, \text{ or } \frac{1}{a} \sin 4\theta > 0. \quad (7, 8)$$

Since a and b may each take either a positive or a negative value, we have four possibilities, viz.,

(I) If a is +ve, then $\sin 4\theta$ is +ve : (a) If b is +ve, then $\tan 4\theta$ is +ve, and consequently $\cos 4\theta$ is +ve ; and therefore θ lies between 0° and $22\frac{1}{2}^\circ$. (b) If b is -ve, then $\tan 4\theta$ is -ve, and consequently $\cos 4\theta$ is -ve ; and therefore θ lies between $22\frac{1}{2}^\circ$ and 45° .

(II) If a is -ve, then $\sin 4\theta$ is -ve : (a) If b is -ve, then $\tan 4\theta$ is +ve, and consequently $\cos 4\theta$ is -ve ; and therefore θ lies between 45° and $67\frac{1}{2}^\circ$. (b) If b is +ve, then $\tan \theta$ is -ve, and consequently $\cos 4\theta$ is +ve ; and therefore θ lies between $67\frac{1}{2}^\circ$ and 90° .

These results are summarized in Table I.

TABLE I. DETERMINATION OF THE ANGLE θ

Sign of Numerator	Sign of Denominator	Angle of Rotation (θ)	Limits for θ
+	+	$\frac{\varphi}{4}$	0° to $22\frac{1}{2}^\circ$
+	-	$\frac{\pi - \varphi}{4}$	$22\frac{1}{2}^\circ$ to 45°
-	-	$\frac{\varphi + \pi}{4}$	45° to $67\frac{1}{2}^\circ$
-	+	$\frac{2\pi - \varphi}{4}$	$67\frac{1}{2}^\circ$ to 90°

Note. $\varphi = \arctan 4\theta = \arctan \frac{a}{b}$.

The procedure of solving for θ and making the transformation for a pair of factors at a time is well suited to punched-card or electronic computer equipment. Either unities or communalities may be inserted in the leading diagonal of the correlation matrix ; any initial factoring may be used ; and the solution may be calculated for any desired number of factors.¹

¹ A detailed description of the full computational procedure can be obtained from the authors upon request.

IV. COMPARISON WITH CARROLL'S ANALYTIC PROCEDURE

Our formulation of the rotational problem¹ results in a procedure which is equivalent mathematically to the one proposed by Carroll (3). This requires the minimization of the cross-products of the columns of squared loadings $\sum_i \sum_{j < k} f'_{ij}{}^2 f'_{ik}{}^2$ (to use our notation rather than Carroll's).

In the i th row of any factoring F , $\sum_j f_{ij}{}^2 = r_{ii}$, a diagonal element in the correlation matrix: it follows that

$$(\sum_j f_{ij}{}^2)^2 = \sum_j f_{ij}{}^4 + 2 \sum_{j < k} f_{ij}{}^2 f_{ik}{}^2 = r_{ii}{}^2. \quad (9)$$

Summing over the n rows, we obtain

$$\sum_i (\sum_j f_{ij}{}^2)^2 = \sum_i \sum_j f_{ij}{}^4 + 2 \sum_i \sum_{j < k} (f_{ij}{}^2 \cdot f_{ik}{}^2) = \text{const.}, \quad (10)$$

i.e., the total test variance accounted for is constant over any orthogonal transformation. Then, if $\sum_i \sum_j f_{ij}{}^4$ is maximized, it follows that $\sum_i \sum_{j < k} (f_{ij}{}^2 \cdot f_{ik}{}^2)$ is minimized.

Carroll (3, p. 24) aimed at finding a mathematical formulation for criteria 3, 4 and 5 in Thurstone's requirements (6, p. 335) for simple structure. Our approach was directed towards the concentration of test variance, and would seem related rather to criteria 1 and 2 (a large number of zero and near-zero loadings). Thus, although these two procedures were developed to satisfy different criteria, they prove

¹ The method reported here provides an interesting case of several investigators working independently towards an analytic rotational solution, and arriving at rather similar results without knowledge of the findings of the others. The references are: Carroll, J. B. (3). Ferguson, G. A. 'The concept of parsimony in factor analysis.' (Unpublished manuscript, July, 1953.) Saunders, D. R. 'An analytic method for rotation to orthogonal simple structure.' (Princeton, New Jersey: Educational Testing Service Research Bulletin 53-10, August, 1953: see also *Amer. Psychologist*, August, 1953, 8, 428.) Our own paper was first issued in August, 1953.

It may be useful to summarize here some points of resemblance and difference between our approach and the results of Ferguson and of Saunders.

1. *The criterion for rotation.* Ferguson's paper is principally concerned with the rational basis for a parsimonious rotational solution; the logical considerations are examined more fully in his than in any of the other papers. He has suggested taking the sum of squares of products of co-ordinates as a measure of parsimony, and shown this to be equivalent to the maximization of fourth powers. Saunders has proposed maximizing the kurtosis of the factor loadings; this also leads to the fourth-power function for maximization in the orthogonal case.

2. *The computational procedure.* Saunders' mathematical development and computational procedure differ from those presented here. The method described above appears simpler for punched-card and desk calculator work; but Saunders' function, because of the numerator and the denominator being in the same form, is more easily programmed for an electronic computer. We have used Saunders' method in our machine orders for the Illiac computer. Wrigley, Saunders, and Neuhaus have since collaborated in preparing a paper entitled 'The application of an analytical method of rotation to Thurstone's primary mental abilities test battery,' presented at the May 1954 meeting of the Midwestern Psychological Association, Columbus, Ohio (Mimeographed; available on request). Ferguson presented no computational procedure.

Professor Burt has also sent us an alternative proof based on the principle that the kind of factors required will be those that maximize the numerical *difference* between the columns of factor saturations. To abolish the signs of the saturations each saturation is first squared; to abolish the signs of the differences, the differences themselves are then squared: so that (in the notation used above) the quantity to be maximized is, for any pair of factors, i and j ,

$$\sum_k (f'_{ki}{}^2 - f'_{kj}{}^2)^2.$$

As he points out, this leads to an expression which in form is very similar to that obtained in finding the principal axes for the normal correlation surface (Yule and Kendall, *Introduction to Statistics*, p. 231, eqn. 12.10).

to be mathematically equivalent. The advantage of formulating the problem in terms of maximizing fourth powers is that it leads to a simpler solution and easier computational procedure.

The algebraic identity of Carroll's procedure and the quartimax holds only for orthogonal solutions. Carroll's procedure is applicable to oblique transformation as well. Extension of the quartimax principle to the oblique case will involve finding the oblique transformation matrix U such that

$$mn \sum_i \sum_j f_{ij}'^4 - [\sum_i \sum_j f_{ij}'^2]^2 = \max. \quad (11)$$

This might be possible either by directly solving a set of non-linear equations of the third degree, or by an iterative procedure involving rotation of the i th axis successively with respect to each of the planes formed by the i th axis and all other axes, repeating the procedure with the $(i + 1)$ th axis, and so on.¹

V. NUMERICAL EXAMPLES

Three examples will be presented : (a) the Burt-Pearson matrix of seven physical measurements, (b) the Holzinger and Harman matrix of 24 psychological tests, and (c) the Thurstone box problem (full 20-variable matrix and reduced 11-variable matrix). These have been calculated on the Illiac (University of Illinois electronic computer). With this a full cycle of operations consists of $m(m - 1)/2$ rotations in the following order : 1, 2 ; 1, 3 ; . . . ; 1, m ; 2, 3 ; . . . ; 2, m ; . . . ; $m - 1$, m , with as many cycles as are needed for convergence. In these examples, calculations were continued until the convergence of the sum of fourth powers to the eighth decimal place. Five cycles were needed for the Holzinger and Harman example, and three for each of the others. Time for the input of orders and data, the Illiac calculation, and the punching of results on tape ranged from about 20 seconds for the Burt-Pearson example to about a minute for the Holzinger and Harman example.

TABLE II. BURT-PEARSON MATRIX: COMPARISON OF QUARTIMAX SOLUTION AND BURT'S GROUP FACTOR ANALYSIS

Body Measurement	Quartimax Factors		Burt's Overlapping Group Factors		
	1	2	1	2	3
Head length ..	.329	.427	.420	.019	.337
Head breadth ..	.124	.861	.334	-.090	.802
Face breadth ..	.319	.679	.479	.071	.577
Foot	.846	.130	.686	.513	.036
Forearm	.972	-.011	.595	.788	-.002
Height	.852	.110	.836	.353	-.094
Finger	.876	.047	.542	.709	.061
Square-Sum ..	3.379	1.416	2.334	1.525	1.104

The Burt-Pearson matrix. Table II gives the quartimax solution for the first two weighted summation (principal axes) factors for the Burt-Pearson correlation matrix (1, p. 116). Burt's group factor results are presented alongside. The separation of head measurements from limb measurements is found, but Burt's method gives in addition a 'basic' factor of

¹ We owe this latter suggestion to Kern W. Dickman.

general bodily size. This basic factor does not appear in the quartimax solution, and loadings on the quartimax factors are somewhat higher than on Burt's two group factors. The quartimax results are thus closer to a simple structure solution.

The Holzinger and Harman matrix. Table III gives the quartimax rotation for Holzinger and Harman's 24-variable centroid solution (4, p. 191), and alongside the results for Holzinger and Harman's orthogonal graphical rotation (4, p. 232). The two sets of results agree fairly well. The principal change is that the verbal factor is somewhat larger in the quartimax solution. Table IV shows rate of convergence for this example. By the end of two cycles, convergence is probably adequate from a practical standpoint.

TABLE III. HOLZINGER AND HARMAN TABLE FOR 24 PSYCHOLOGICAL TESTS : COMPARISON OF QUARTIMAX SOLUTION AND ORTHOGONAL ROTATED SOLUTION

Test	Quartimax Factors				Orthogonal Rotated Factors			
	1	2	3	4	Verbal	Speed	Deduction	Memory
1	.369	.190	.599	.068	.098	.324	.616	.203
2	.245	.066	.384	.039	.068	.153	.410	.130
3	.313	.010	.475	.013	.095	.119	.532	.131
4	.359	.070	.463	—012	.150	.177	.530	.115
5	.806	.135	—016	—039	.746	.150	.255	.154
6	.812	.030	.000	.055	.722	.051	.279	.251
7	.854	.072	—044	—095	.807	.085	.271	.111
8	.660	.202	.197	—034	.535	.258	.379	.140
9	.857	—058	—020	.098	.757	—039	.287	.305
10	.234	.700	—124	.112	.277	.664	—191	.139
11	.310	.616	.011	.231	.267	.612	—045	.286
12	.164	.688	.191	—012	.128	.717	.091	.029
13	.350	.569	.323	—082	.239	.634	.312	.021
14	.324	.190	—026	.424	.231	.188	—019	.482
15	.251	.114	.101	.448	.106	.138	.078	.500
16	.289	.134	.372	.368	.053	.218	.345	.454
17	.285	.239	.022	.569	.153	.245	—027	.616
18	.215	.324	.302	.467	.006	.386	.199	.522
19	.276	.180	.189	.323	.123	.223	.175	.392
20	.518	.090	.354	.142	.307	.177	.459	.292
21	.347	.385	.347	.144	.172	.460	.331	.245
22	.526	.041	.295	.255	.311	.117	.400	.399
23	.546	.187	.441	.087	.314	.291	.537	.251
24	.487	.432	.101	.196	.387	.457	.142	.306
Square-Sum	5.587	2.422	1.958	1.418	3.430	2.917	2.683	2.358

TABLE IV. HOLZINGER AND HARMAN TABLE : RATE OF CONVERGENCE AND SUMS OF FOURTH POWERS WITH QUARTIMAX SOLUTION

Cycle	Sum of Fourth Powers	Increase in Sum from Previous Cycle	Percentage of Total Increase
0	3.06391674		
1	3.92477171	86085497	96.097
2	3.95886327	3409156	99.903
3	3.95972451	86124	99.999
4	3.95973460	1009	100.000*
5	3.95973487	27	
6	3.95973487	0	

* Figures rounded to three decimal places.

The Thurstone box problem. Quartimax solutions are given for the full 20-variable example (6, p. 136) in Table V, and the reduced 11-variable example (6, p. 374) in Table VI. Both are presented because of the contrasting results. In the 20-variable example, the quartimax solution agrees rather closely with Thurstone's, though the one is orthogonal and the other oblique. In the 11-variable example, there is less agreement. 'Depth' measurements appear in a separate quartimax factor, but 'length' measurements are not effectively separated from 'breadth' measurements.

TABLE V. THURSTONE 20-VARIABLE BOX PROBLEM : COMPARISON OF QUARTIMAX SOLUTION AND OBLIQUE SIMPLE STRUCTURE SOLUTION

Variable	Quartimax Factors			Thurstone's Oblique Factors		
	1	2	3	X	Y	Z
x^2	.992	.087	.060	.969	.003	-.003
y^2	.161	.966	.171	.025	.939	.005
z^2	.046	.076	.986	.007	.023	.957
xy	.592	.787	.159	.476	.727	.000
xz	.428	.085	.892	.387	.005	.844
yz	.098	.447	.884	.012	.393	.794
$\sqrt{x^2 + y^2}$	.808	.586	.111	.719	.512	-.025
$\sqrt{x^2 + z^2}$	.887	.111	.452	.850	.016	.384
$\sqrt{y^2 + z^2}$	.115	.803	.588	-.011	.760	.444
$2x + 2y$	.719	.692	.141	.615	.623	-.008
$2x + 2z$	.709	.101	.699	.668	.008	.638
$2y + 2z$	.105	.659	.751	-.007	.610	.628
$\log x$	.983	.118	.055	.956	.034	-.013
$\log y$	.111	.946	.225	-.024	.921	.063
$\log z$	.061	.034	.959	.028	-.019	.937
xyz	.367	.434	.804	.282	.362	.704
$\sqrt{x^2 + y^2 + z^2}$	.735	.543	.402	.644	.460	.273
e^x	.964	.050	.064	.947	-.032	.008
e^y	.203	.934	.123	.072	.906	-.040
e^z	.045	.085	.978	.004	.032	.948
Square-Sum	6.669	5.972	6.999	5.692	5.181	5.758

TABLE VI. THURSTONE 11-VARIABLE BOX PROBLEM: COMPARISON OF QUARTIMAX SOLUTION AND OBLIQUE SIMPLE STRUCTURE SOLUTION

Variable	Quartimax Factors			Thurstone's Oblique Factors		
	1	2	3	Y	Z	X
<i>x</i>	·684	·047	·651	·05	·00	·90
<i>y</i>	·834	—·004	—·479	·88	·01	·04
<i>z</i>	·416	·794	—·075	·05	·79	·03
<i>xy</i>	·992	—·037	·115	·63	—·06	·62
<i>yz</i>	·714	·562	—·401	·54	·57	—·05
<i>x²y</i>	·919	·028	·412	·37	—·01	·82
<i>xy²</i>	·956	·029	—·177	·76	·02	·35
<i>2x + 2y</i>	·997	—·071	—·001	·71	—·09	·53
$\sqrt{x^2 + y^2}$	·985	—·066	—·014	·71	—·08	·52
$\sqrt{x^2 + z^2}$	·592	·684	·381	—·07	·65	·52
<i>xyz</i>	·882	·445	·035	·43	·43	·42
Square-Sum	7·681	1·627	1·181	3·380	1·575	2·992

VI. DISCUSSION

The search for simple structure is usually very laborious, and, since much of the work consists in the plotting of graphs, the technique is not readily adapted to mechanized computation.¹ For any psychologist with access to an electronic computer or punched-card equipment, the quartimax method enormously reduces the time and labour required. For example, each cycle of 78 transformations for Thurstone's 13-factor primary mental abilities is calculated on Illiac in about $1\frac{1}{2}$ minutes. The sum of fourth powers converges (to eight decimal places) after 27 cycles—about 45 minutes in all, allowing for input and output time. The main mental abilities were clearly revealed in the quartimax solution.² More than 99 per cent. of the increase in the sum of fourth powers occurred in the first four cycles, so that a reasonable solution might have been obtained in about ten minutes' machine time.

Thus the psychologist with good machine aids who is seeking simple structure (or some other form of group factor analysis) will find it computationally economical to start by calculating a quartimax solution. This will generally reveal the group structuring of the variables with sufficient clarity to make the graphical rotation much more rapid than if an unrotated centroid (or principal axes) solution provided the starting point. These are our grounds for referring to the quartimax method as "an approach to orthogonal simple structure."

A more fundamental problem is whether the quartimax method gives results that are meaningful in and for themselves. Their characteristics may perhaps be most succinctly expressed by considering resemblances to and differences from simple structure. Let us take the five criteria which Thurstone (6, p. 335) suggested for deciding when simple structure has been reached. We find that the second criterion of Thurstone (that there should be some zeros in each column of the factor matrix) is the one which the quartimax method is least certain to meet. It will therefore be convenient to consider the other criteria first.

Criterion 1. Each row of the factor loading matrix should have at least one zero. The quartimax method implies no stipulation as to the number of zeros (or near-zeros) appearing in each row, nor are tests which fail to meet this criterion eliminated from the battery. It can, however, probably be claimed that each test will be expressed in terms of the minimum number of factors, consonant with the demands of an orthogonal transformation. (Oblique transformations may increase the number of zeros in the factor matrix, thereby reducing the number of factors on which any test is loaded.) The quartimax method has one advantage

¹ It should be mentioned, however, that an electronic factor rotator which has been constructed and is in regular operation at the Personnel Research Section, Office of the Adjutant General, Washington, D.C., uses the standard graphical procedure.

² Cf. Wrigley, Saunders, and Neuhaus, *op. cit. sup.*

over the simple structure criterion. A calculated factor loading will rarely if ever be exactly zero. In trying to get simple structure, the standard practice seems to be for the investigator to regard all loadings within the range $\pm x$ as essentially equivalent to zero because of sampling error. Lacking a satisfactory test of significance, the investigator has himself to define the limits $\pm x$ and a second investigator might well choose different limits and reject the structure attained by the first. The quartimax method does not need to define the limits, but reduces the low entries as much as possible.

Criteria 3-5. For every pair of columns there should be (a) several tests with zeros in one column but not the other, (b) a large proportion with zeros in both columns, (c) only a small number of tests with non-vanishing entries in both columns. The maximization of fourth powers has been shown to lead to the minimization of the cross-products of the columns of squared factor loadings; and this, in our view and in Carroll's, implies that criteria 3-5 are reasonably fulfilled.

Criterion 2. For each column there shall be m tests whose factor loadings are zero (m = no. of factors). This excludes any general factor and ensures that all factors are of approximately the same size. In a quartimax solution, however, a general factor sometimes appears, and factors may differ greatly in size: the first factor for the 11-variable box problem, for example, has no loading less than .4. Sometimes one of the group factors extends much further than in a graphical solution, and so approximates to a general factor. The first factor in the Holzinger and Harman example has only one loading less than .2. It would be interesting to know why a general factor appears in the 11-variable box problem, but not in the 20-variable box problem. In Thurstone's figures, the former differs from the latter in at least two respects: (a) the first centroid factor is larger in the former (76.3 per cent. of the retained variance as against 67.4 per cent.), and (b) simple structure factors are more oblique, perhaps because of the larger first factor (an average correlation of primary factors of .27 as against .19). A large first principal axes (or centroid) factor implies greater inequalities in the initial factor loadings, and probably means less extensive quartimax transformation. It seems a reasonable feature of the method that the stronger the first factor is in the principal axes solution, the more likely is it that a general factor will appear in the rotated quartimax solution.

We may summarize this discussion as follows. Thurstone presents certain formal requirements. They will sometimes be satisfied by an orthogonal transformation, but more frequently an oblique transformation will be required. As these criteria are applied in general to all factor analyses, there will be occasions in which simple structure can be easily attained, and others in which it can only be reached with great difficulty or not at all.¹ Further, some arbitrary decision still seems necessary as to the boundaries of the hyperplanes—i.e., as to the size of loading that is to be regarded as not significantly different from zero. There is some evidence that the difficulty of finding simple structure is influenced by (a) the number of factors extracted, (b) the average size of correlation, and (c) the number of variables. The quartimax method has only a single requirement, viz., that the variance of the squared factor loadings be maximized. The requirements about the definite number of zeros to be present in each row and column have been eliminated. Hence we can no longer say whether there will be a general factor.

Thurstone contends that group factors are to be preferred to general factors on psychological grounds. On the other hand, Burt sponsors a Three-Factor Theory, and insists that we must allow for the possibility of three types of factor—general, group, and specific.² Here the quartimax method approaches more closely to Burt's position than to Thurstone's: any of the three types of factor may emerge.

¹ It is well known that orthogonal simple structure can be attained only under certain specifiable conditions: see Ledermann (5).

² These three types of factor also appear in Holzinger's bifactor method. The remarks made in this and the next four paragraphs apply for the most part to Holzinger's bifactor method as well as to Burt's method of group factor analysis.

There seem to be three principal differences between the quartimax method and Burt's method of group factor analysis (2). First, the former does not imply any sharp distinction between general and group factors. The difference becomes one of degree rather than of kind. Factors may often be secured which cannot be classified unequivocally as either. But this difference must not be stressed too greatly, since Burt himself has a place in his system for 'overlapping group factors.' Secondly, Burt's procedure depends on a prior classification of variables into groups, according to the sign-pattern of the initial factor loadings. No such classification is needed in the quartimax method. The classification emerges as a result of the procedure. Thirdly, Burt's method sometimes introduces more factors than the quartimax. Consider the results of Table II. Only two quartimax factors are needed to distinguish between head measurements and limb measurements. It operates upon the initial set of factors without increasing their number. Burt has three—a basic factor followed by two group factors. His classification by sign-pattern means that the number of factors in his group factor analysis may be somewhat larger than the number of initial factors. Mathematically the quartimax method is the more parsimonious. On the other hand, Burt's solution conforms more to the generally accepted view that special factors for growth of head and of limbs are superimposed upon a factor for general growth. The dilemma that faces us in deciding upon the appropriate number of factors is that there is as yet no general agreement about the purpose of rotation or about the type of rotational solution most useful to the psychologist. Until these issues are decided, there can be no universally accepted rotational procedure.

The quartimax method has been presented as an illustration of the practicability of developing an analytic procedure suitable for high-speed machines. Its characteristic feature, as contrasted both with Burt's group factor analysis and with Thurstone's simple structure, is its greater flexibility. No prior classification of variables is needed, and the test battery does not have to be 'purified.' There is no assumption about the presence or absence of a general (or 'basic') factor. By accepting a general factor whenever it appears, we avoid the additional complications resulting from the use of oblique axes and the re-emergence of the general factor as a second-order factor. This greater flexibility is convenient computationally; but the method will have to be tried far more extensively before we can decide whether this greater flexibility is advantageous or disadvantageous from a psychological standpoint.

VII. SUMMARY

An analytic method, named 'quartimax,' for the orthogonal rotation of factor loadings has been proposed. It consists in maximizing the sum of fourth powers of factor loadings. It is objective (i.e., non-judgmental), and adaptable to high-speed computational devices. The relations of this procedure to Thurstone's 'simple structure' and Burt's method of group factor analysis have been discussed, and three examples presented.

REFERENCES

1. Burt, C. (1949). 'Alternative methods of factor analysis and their relations to Pearson's method of principal axes.' *Brit. J. Psychol., Stat. Sect.*, II, 98–121.
2. Burt, C. (1950). 'Group factor analysis.' *Brit. J. Psychol., Stat. Sect.*, III, 40–75.
3. Carroll, J. B. (1953). 'An analytical solution for approximating simple structure in factor analysis.' *Psychometrika*, XVIII, 23–38.
4. Holzinger, K. J., and Harman, H. H. (1941). *Factor Analysis*. Chicago: University of Chicago Press.
5. Ledermann, W. (1936). 'Some mathematical remarks concerning boundary conditions in the factorial analysis of ability.' *Psychometrika*, I, 165–74.
6. Thurstone, L. L. (1947). *Multiple-factor Analysis*. Chicago: University of Chicago Press.

A TEST OF SIGNIFICANCE FOR THE
RELATION BETWEEN *m* RANKINGS
AND *k* RANKED CATEGORIES

By A. R. JONCKHEERE

Department of Psychology, University College, London

I. *Problem and Illustrations.* II. *The Statistic Proposed.* III. *The Sampling Distribution of P.*
IV. *The Distribution for Large Samples.* V. *Examples.* VI. *Summary.*

I. PROBLEM AND ILLUSTRATIONS

The following situation often arises during investigations in educational and social psychology : *n* objects are ranked for some characteristic by each of *m* judges ; the investigator regards these *n* objects as grouped into categories, and wishes to determine whether the *m* sets of rankings obtained from the judges agree with a rank order predicted for these categories. In a recent contribution to this *Journal* (6), Whitfield described a technique suitable for such a problem when only two categories are involved and when one of them contains only a single object. The present paper attempts to generalize his method, and so to increase the variety of situations to which it may be applied.

Let *n* objects be divided into *k* categories such that the *r*th category contains *l_r* objects, where $\sum_{r=1}^k l_r = n$. The categories will be considered as ranked in order of increasing value ; and hence the objects within them may be allotted appropriately tied ranks. Objects in the *r*th category will thus receive a rank value of $\frac{1}{2}(l_r + 1) + \sum_{i=1}^{r-1} l_i$. The experimenter wishes to test the null hypothesis that full rankings of the *n* objects obtained from *m* judges do not differ substantially from those to be expected had the ranks been allotted at random, and, if they do, to determine whether any departure from such randomness is in agreement with the predicted rank order for the *k* categories.

Two artificial examples will make clear the applications of this abstract scheme.

Let us suppose that, in a given school, 9 pupils are promoted to class I. Of these, 3 are drawn from class II, 1 from class III, 2 from class IV, and 3 from class V. The headmaster predicts that their ability will diminish from class to class. Assigning tied ranks to the pupils coming from the various classes in accordance with this prediction we have the following result :

<i>Class from which pupils are promoted</i>	V	IV	III	II
<i>Number of pupils</i>	3	2	1	3
<i>Predicted rank</i>	2	4½	6	8

Four teachers who take class I are asked to rank the pupils in order of merit. In the present example, using the symbols proposed above, we should have : *n* = 9, *m* = 4, *k* = 4, *l*₁ = 3, *l*₂ = 2, *l*₃ = 1, *l*₄ = 3.

The problem is to find whether there is a significant agreement between (i) the tied rankings predicted by the headmaster and (ii) the empirical rankings supplied by the four teachers.¹

¹ I am indebted to Sir Cyril Burt for this example : cf. his paper on ‘The relative merits of ranks and paired comparisons,’ this *Journal*, VI, pp. 112–19.

A Test of Significance

A second application may be found in researches on group activities. Suppose a group of 10 people has among its members 6 who have been coached to play special roles in certain group activities. One of them is to behave (say) 'democratically,' 3 'autocratically,' and 2 'anarchically.' At the end of an experimental session, each of the four members who have not been coached ranks his 9 fellow members for some characteristic, such as 'rigidity of behaviour.' The experimenter advances the hypothesis that the 'autocratic,' 'anarchic,' and 'democratic' members will be ranked in that order for decreasing 'rigidity,' with the remaining group members showing the least degree of this characteristic. Making allowance for ties, we have for this example :

Category	Ordinary members	Democratic members	Anarchic members	Autocratic members
Number ranked as below ..	3	1	2	3
Predicted rank	2	4	5½	8

Here, in terms of the symbols proposed, $n = 9$, $m = 4$, $k = 4$, $l_1 = 3$, $l_2 = 1$, $l_3 = 2$, $l_4 = 3$. To test the experimenter's hypothesis we require some means of determining whether the set of four rankings actually obtained differ from those we should expect had they been assigned at random, and whether the differences found are in significant agreement with the predicted ordering of the categories of behaviour.

A similar problem arises in assessing the over-all significance of trend in a set of time series. Suppose that an experimenter predicts that the successive measures forming a time series will show an upward trend. As data for analysis he obtains time series from a number of subjects. His null hypothesis is that no trend is present ; and this he wishes to test against the alternative of an upward trend. If, in terms of the abstract scheme given above, we identify n with the number of successive measures in each of m independent time series, we have a case in which the number of categories used by the experimenter in making his prediction is identical with n . The test here proposed will be appropriate to this situation when $k = n$, and hence all the $l_i = 1$. The measures will then be ranked. An example of this use of the test is given in section V.

The abstract scheme thus outlined could be made still more general by allowing each of the m judges to rank differing numbers of the n objects, or each of the m subjects to produce time series of differing length. The scheme given above, however, appears sufficient for most practical purposes.

II. THE STATISTIC PROPOSED

A measure of agreement between a theoretical ranking and any particular observed ranking obtained from one of the m judges is given by Kendall's S coefficient (4), when allowance is made for the ties occurring in the predicted ranking. The following shows the method of computation of S for a ranking obtained from one judge in the example given above relating to group behaviour. Suppose, for instance, that we obtain the following result from the i th judge :

	Ordinary members	Democratic members	Anarchic members	Autocratic members
Ranking obtained from i th judge ..	2, 4, 5	1	3, 7	6, 8, 9
Experimenter's predicted ranking ..	2, 2, 2	4	5½, 5½	8, 8, 8

To compute the appropriate value of S when allowance has to be made for the ties occurring in the predicted ranking, we proceed as follows : (i) Group the ranking obtained from the judge into the k categories so that the categories are arranged in the predicted sequence of increasing value : in the above example this has already been done. (ii) Taking each of the ranks in turn, count the number in the succeeding groups which are of greater value. Let the total count for all the ranks occurring in the first $k - 1$ categories be P_i . In our example we have

from the first category of ordinary members 5 + 4 + 4 ;
from the second category of 'democratic' members 5 ;
from the third category of 'anarchic' members 3 + 2.

No such count of course can be made from the fourth category of 'autocratic' members. Hence for the over-all total we have $P_i = 23$.

(iii) The value of S may now be obtained from the expression

$$S_i = 2P_i - \sum_{i=1}^{k-1} \sum_{j=1, i}^k l_i l_j . \quad (1)$$

With the above example we have

$$S_i = 2 \cdot 23 - \{ (3 \times 1) + (3 \times 2) + (3 \times 3) + (1 \times 2) + (1 \times 3) + (2 \times 3) \} = 17.$$

For the present test it is more convenient to employ P_i as a statistic ; the computation can then be done mentally.

If $k = 2$ (i.e., with only two categories) a still simpler method of computation is available : we have but to add the ranks allotted to the objects within the second of the two categories. If this sum is represented by X_i , then

$$P_i = X_i - \frac{1}{2}l_2(l_2 + 1). \quad (2)$$

Suppose, for example, that in an experiment on group behaviour there are only two categories of members : 4 'ordinary' and 6 'autocratic.' The ranking obtained from a judge who ranks these ten individuals for their 'rigidity' might be as follows :

	Ordinary members	Autocratic members
Ranking of i th judge	1, 2, 3, 7	4, 5, 6, 8, 9, 10
Experimenter's predicted ranking . .	$2\frac{1}{2}, 2\frac{1}{2}, 2\frac{1}{2}, 2\frac{1}{2}$	$7\frac{1}{2}, 7\frac{1}{2}, 7\frac{1}{2}, 7\frac{1}{2}, 7\frac{1}{2}, 7\frac{1}{2}$

The sum of the ranks allotted to the 'autocratic' members is $X_i = 42$. Hence, employing equation (2), we have $P_i = 42 - \frac{1}{2}6(6 + 1) = 21$.

Up to now we have considered merely a measure (P_i) to be obtained from the i th judge alone. We require, however, an over-all estimate of agreement between theory and fact, based on the results from all m judges. Such an estimate will be obtained if we compute the appropriate values of P_i for each of the m sets of rankings supplied by the judges, and add

them together. Let us denote this sum by $P \equiv \sum_{i=1}^m P_i$.

We may easily construct a measure, with values lying between $+1$ and -1 , to express the agreement between the experimenter's predictions and the m sets of ranks. The extreme values for P will be zero and $m \sum_{i=1}^{k-1} \sum_{j=i+1}^k l_i l_j$. Hence an over-all coefficient of correlation is given by

$$\tau = \frac{2P}{m \sum_{i=1}^{k-1} \sum_{j=i+1}^k l_i l_j} - 1 \quad (3)$$

When $m = 1$ and $k = n$, then $l_i = 1$; and in this case τ becomes identical with Kendall's τ coefficient of correlation between two rankings.

III. THE SAMPLING DISTRIBUTION OF P

The null hypothesis, which the experimenter wishes to reject in favour of his predicted alternative, is tantamount to supposing that all possible permutations of the n ranks produced by the m judges are equally likely. We therefore require the distribution of the P 's obtained from such a population.

In another paper (3), expressions are given for all the cumulants of the distribution of S , when one ranking contains ties of any extent and number. Making the necessary change in the mean and scale, and employing the above expressions, we can now write down the cumulants of the distribution of P_i :

$$\begin{aligned} \kappa_1(P_i) &= \frac{1}{2} \sum_{i=1}^{k-1} \sum_{j=i+1}^k l_i l_j ; & \kappa_{2\alpha+1} &= 0, \\ \kappa_{2\alpha}(P_i) &= (-1)^{\alpha-1} \frac{B_\alpha}{2\alpha} \left\{ \sum_{p=1}^n p^{2\alpha} - \sum_{r=1}^k \sum_{p=1}^{l_r} p^{2\alpha} \right\}, \end{aligned} \quad (4)$$

where $\alpha \geq 1$; B_α is the α th Bernoulli number ($B_1 = \frac{1}{6}$, $B_2 = \frac{1}{30}$, etc.).

The ranking given by any one judge is assumed to be independent of that given by another. Consequently each of the m individual P_i 's must be distributed independently of the others ; and, using a well-known result for the cumulants of the distribution of a sum

A Test of Significance

of independent random variables,¹ the cumulants of P can therefore be expressed as follows :

$$\begin{aligned} \kappa_1(P) &= \frac{1}{2}m \sum_{i=1}^{k-1} \sum_{j=1+i}^k l_i l_j; & \kappa_{2\alpha+1}(P) &= 0, \\ \kappa_{2\alpha}(P) &= m(-)^{\alpha-1} \frac{B_\alpha}{2\alpha} \left\{ \sum_{p=1}^n p^{2\alpha} - \sum_{r=1}^k l_r \sum_{p=1}^r p^{2\alpha} \right\}. \end{aligned} \quad (5)$$

α and B_α having the same significance as in the previous expressions for the distribution of P_i . In particular, expanding the above expressions for $\alpha = 1, 2$ we have :

$$\begin{aligned} \kappa_2(P) &= \frac{m}{72} \left\{ n^3(2n+3) - \sum_{r=1}^k l_r^2(2l_r+3) \right\}, \\ \kappa_3(P) &= 0, \\ \kappa_4(P) &= -\frac{m}{3600} \left\{ n^3(6n^2+15n+10) - \sum_{r=1}^k l_r^3(6l_r^2+15l_r+10) \right\}, \end{aligned}$$

and hence $\gamma_1(P) = 0$

$$\gamma_2(P) = -\frac{36}{25} \frac{\left\{ n^3(6n^2+15n+10) - \sum_{r=1}^k l_r^3(6l_r^2+15l_r+10) \right\}}{m \left[n^2(2n+3) - \sum_{r=1}^k l_r^2(2l_r+3) \right]^2}. \quad (6)$$

If we put $k = 2$, we may employ the simplified computation, mentioned above, in terms of the sums of the ranks. In this case the grand total sum of ranks allotted to objects within the second category, denoted by X , is related to the total P by the expression

$$X \equiv \sum_{i=1}^m x_i = P + \frac{1}{2}ml_2(l_2+1).$$

Hence for this case the cumulants of the grand total P are given by

$$\begin{aligned} \kappa_1(X) &= \frac{1}{2}ml_2(n+1) \\ \kappa_2(X) &= \frac{1}{12}ml_1l_2(n+1) \\ \kappa_{2\alpha}(X) &= m(-)^{\alpha-1} \frac{B_\alpha}{2\alpha} \left\{ \sum_{p=1}^n p^{2\alpha} - \sum_{p=1}^{l_1} p^{2\alpha} - \sum_{p=1}^{l_2} p^{2\alpha} \right\}. \end{aligned} \quad (7)$$

These expressions were obtained by Haldane and Smith (2) in discussing a test for birth order effects. Moreover, if we put $l_1 = n-1$ and $l_2 = 1$ in the above expressions for the case when $k = 2$, we obtain the results given by Stuart (5), and, for the values $\alpha = 1, 2$, the moments obtained by Whitfield (6).

IV. THE DISTRIBUTION FOR LARGE SAMPLES

It has been shown elsewhere that, if *more than one* of the ties increases in range without limit, then the distribution of S , when only one of the two sets of ranks contains ties, will be normal. It follows that, with this condition (more than one of the l_i 's increasing without limit), P_i , and hence P , is normally distributed, regardless of the size of m .

If only *one* of the ties (or l_i 's) increases without limit, then the distribution of S when ties occur in only one array of ranks is not normal. It follows that the distribution of a single P_i is non-normal. Hence P , a finite sum of the P_i 's, will not be normally distributed. In this case normality will only be attained if m is allowed to increase without limit—precisely the situation occurring with the test discussed by Whitfield and Stuart, where $l_1 = n-1$ and $l_2 = 1$. For as n tends to infinity only one group increases in size.

In general the tendency towards normality is rapid, as can be seen from an examination of the value of $\gamma_2(P)$ for the general case. Hence we may test the significance of a given P by transforming it into a unit normal deviate and referring to tables of the normal distribution. Since the interval between all possible adjacent values of P is always unity, an improvement in the approximation to the true distribution will be obtained if $\frac{1}{2}$ is subtracted from an obtained value of P prior to its reduction to unit normal form.

¹ See, for example, H. Cramér (1), p. 192.

We have finally, as a test for the significance of a given P , that

$$z = \frac{P - \frac{1}{2} - \frac{m}{2} \sum_{i=1}^{k-1} \sum_{j=1+i}^k l_i l_j}{\left\{ \frac{m}{72} [n^2(2n+3) - \sum_{r=1}^k l_r^2(2l_r+3)] \right\}^{\frac{1}{2}}} \quad (8)$$

is distributed, for large m or l_i , as a unit normal deviate.¹ Doubts about the adequacy of the normal approximation may be settled by computing the appropriate value of $\gamma_2(P)$. The closer this value is to zero, the more one may have confidence that the approximation is adequate. Further, since the value of $\gamma_2(P)$ is negative, the distribution will tend to be platykurtic; and hence, taking z from the tables of the normal distribution, we shall have, for low probabilities, estimates which are on the conservative side.

V. EXAMPLES

(a) Suppose that in the experiment with the social groups described above the rankings obtained from four judges are those shown in Table I.

TABLE I. GROUP EXPERIMENT: FOUR CATEGORIES

Judge	Ordinary member	Rank given to—			Value of P_i
		Democratic member	Anarchic member	Autocratic member	
i	2, 4, 5,	1,	3, 7,	6, 8, 9	23
ii	1, 6, 7,	2,	3, 4,	5, 8, 9	21
iii	1, 2, 5,	6,	4, 8,	3, 7, 9	22
iv	2, 3, 5,	1,	6, 7,	4, 8, 9	23
Total (P)					89

Here we have, $n = 9$, $m = 4$, $k = 4$, $l_1 = 3$, $l_2 = 1$, $l_3 = 2$, $l_4 = 3$.

Since $\sum_{i=1}^{k-1} \sum_{j=1+i}^k l_i l_j = 3 \times 1 + 3 \times 2 + 3 \times 3 + 1 \times 2 + 1 \times 3 + 2 \times 3 = 29$, the over-all agreement between the experimenter's prediction and the obtained rankings is given by

$$\tau = \frac{2P}{m \sum_{i=1}^{k-1} \sum_{j=1+i}^k l_i l_j} - 1 = \frac{2 \times 89}{4 \times 29} - 1 = 0.53.$$

The expected mean value of P on the null hypothesis is

$$\kappa_1(P) = \frac{m}{2} \sum_{i=1}^{k-1} \sum_{j=1+i}^k l_i l_j = \frac{4}{2} \times 29 = 58,$$

while the variance is

$$\begin{aligned} \kappa_2(P) &= \frac{m}{72} \left\{ n^2(2n+3) - \sum_{r=1}^k l_r^2(2l_r+3) \right\} \\ &= \frac{4}{72} [9^2(2 \times 9 + 3) - 3^2(2 \times 3 + 3) - 1^2(2 \times 1 + 3) - 2^2(2 \times 2 + 3) - 3^2(2 \times 3 + 3)] \\ &= 251/3. \end{aligned}$$

Hence

$$z = \frac{P - \frac{1}{2} - \kappa_1(P)}{\sqrt{\kappa_2(P)}} = \frac{89 - \frac{1}{2} - 58}{\sqrt{\frac{251}{3}}} = 3.33.$$

¹ For estimating significance a 'one-tailed' test should be used.

A Test of Significance

For the values of n and l_i 's in this example, $\gamma_2 = -072$, indicating that the normal approximation is adequate.

We conclude that the experimenter may reject the hypothesis of randomness with considerable confidence, and accept the alternative hypothesis of an ordered sequence of increasing 'rigidity' in the several social categories.

(b) A further example will illustrate the computation required when only two different categories are under consideration. Suppose that an experiment yields the data given in Table II, where the hypothesis alternative to randomness is that 'ordinary' members are less 'dogmatic' than 'autocratic' members.

TABLE II. GROUP EXPERIMENT 2: TWO CATEGORIES

Judge	Rank given to—		Value of x_i
	Ordinary members	Autocratic members	
i	1, 2, 3, 7,	4, 5, 6, 8, 9, 10	42
ii	1, 2, 6, 10,	3, 4, 5, 7, 8, 9	36
iii	1, 2, 3, 10,	4, 5, 6, 7, 8, 9	39
iv	1, 3, 4, 9,	2, 5, 6, 7, 8, 10	38
	Total (X)		155

Here we have $n = 10$, $k = 2$, $m = 4$, $l_1 = 4$, $l_2 = 6$.

The over-all agreement between the experimenter's prediction and the obtained rankings is given, when $k = 2$, by

$$\tau = \frac{2X - ml_2(l_2 + 1)}{ml_1l_2} - 1 = \frac{2 \times 155 - 4 \times 6 \times 7}{4 \times 4 \times 6} = .48.$$

With the null hypothesis the expected mean value of X is given by

$$x_1(X) = \frac{ml_2(n + 1)}{2} = 132,$$

while the variance is

$$x_2(X) = \frac{ml_1l_2(n + 1)}{12} = 88.$$

Hence a unit normal deviate is given by

$$z = \frac{X - \frac{1}{2} - x_1(X)}{\sqrt{x_2(X)}} = \frac{155 - \frac{1}{2} - 132}{\sqrt{88}} = 2.505.$$

With the values for m , l_i , and k assumed in this example we have $\gamma_2 = -098$, a figure which indicates a good approximation to the normal distribution. Hence we may reject the hypothesis of randomness in the allotment of ranks, and accept the alternative hypothesis that 'autocratic' members are more 'dogmatic' than ordinary members.

(c) When employing the test for the assessment of trend, it is unnecessary to rank the scores for the time series before computing the value of P . P may be obtained by simply counting the appropriate inequalities between the scores, as if they were ranks.

As an example, suppose that each of 5 subjects performs 7 trials with a maze: the experimenter wishes to see whether there is any evidence of learning. Considering the number of errors made on each trial, the null hypothesis to be tested is that there is no trend; and the alternative hypothesis is that there is an upward trend in the number of errors as one proceeds from trial 7 to trial 1. The data obtained from such an experiment might be as shown in Table III. Here we have $n = 7$, $m = 5$, $k = 7$, and $l_i = 1$. An indication of the

TABLE III. LEARNING EXPERIMENT : TIME SERIES

Subject	Number of errors made in trial							Value of P_i
	7	6	5	4	3	2	1	
i	0	4	10	8	5	3	12	14
ii	2	0	7	6	8	4	5	13
iii	1	2	9	8	5	4	10	15
iv	0	7	5	10	6	11	9	16
v	4	3	8	2	9	5	7	13
Total	7	16	39	34	33	27	43	71

degree to which the time series shows an over-all upward trend is given by putting $k = n$ in (3).

We have then

$$\tau = \frac{4P}{mn(n-1)} - 1 = \frac{4.71}{5.7.6} - 1 = .35.$$

The expected mean and variance can be obtained by putting $k = n$ and $l = 1$ in the appropriate expressions of (5) and (6). We then have

$$\kappa_1(P) = \frac{1}{2}mn(n-1) = \frac{1}{2}5.7.6 = 52.50,$$

$$\kappa_2(P) = \frac{mn}{72}(n-1)(2n+5) = \frac{5.7.6.19}{72} = 55.42.$$

Hence a unit normal deviate is given by

$$z = \frac{P - \frac{1}{2} - \frac{1}{2}m.n(n-1)}{\left\{\frac{mn}{72}(n-1)(2n+5)\right\}^{\frac{1}{2}}} = \frac{71 - 53}{(55.42)^{\frac{1}{2}}} = 2.42.$$

With the values for n , m , k , and l , assumed in this example we have, from equation (6), $\gamma_2 = -.063$, a figure which indicates a good approximation to the normal distribution. Hence the experimenter can reject the null hypothesis at the 1 per cent. level of significance in favour of the alternative that an upward trend is actually present. If the errors for each trial were added as shown in the bottom row of Table III, and if the appropriate tau-correlation was computed from the order in which these totals lie, we should obtain a value significant at the 15 per cent. level only; hence the null hypothesis of no trend would be accepted. The reason for the discrepancy between the two significance levels is that with the latter no allowance is made for the number of time series involved in the totals.

VI. SUMMARY

1. A technique has been offered for testing the significance of the agreement between a predicted grading of n objects into k ordered categories and rankings of the n objects by m judges. This technique should be of service in both educational and social psychology.

2. A statistic P has been proposed, related to Kendall's S statistic. It consists of the sum of the S values computed between each of the m rankings and the predicted grading, which is itself treated as a rank order with extensive ties.

3. The cumulants of the sampling distribution of P have been derived, and the distribution of the standardized variate is shown to tend to normality for large n in nearly all cases. Attention has been drawn to the exceptions; and three worked examples have been given.

REFERENCES

1. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton : University Press.
2. Haldane, J. B. S., and Smith, C. A. B. (1948). 'A simple exact test for birth-order effect.' *Ann. Eugen.*, XIV, 117-24.
3. Jonckheere, A. R. (1954). 'A distribution-free k -sample test against ordered alternatives.' *Biometrika*, XLI (in the press).
4. Kendall, M. G. (1948). *Rank Correlation Methods*. London : Griffin.
5. Stuart, A. (1954). 'Limit distributions for total rank values.' *Brit. J. Stat. Psychol.*, VII, 50-51.
6. Whitfield, J. W. (1953). 'The distribution of total rank value for one particular object in m rankings of n objects.' *Brit. J. Stat. Psychol.*, VI, 35-40.

THE FULL FACTOR ANALYSIS

By F. W. WARBURTON

Department of Education, University of Manchester

I. *Introduction.* II. *Objections to Specific Factors : (a) Mathematical ; (b) Psychological.* III. *Advantages of the Full Analysis.* IV. *Disadvantages of the Full Analysis.* V. *The Enquiry.* VI. *The Triangular Method of Analysis.* VII. *Results.* VIII. *Summary.*

I. INTRODUCTION

There has recently been a revival of interest in what may be called 'full factor analysis,' the procedure whereby as many factors are extracted as there are tests. This type of analysis is, of course, not new. It is the method favoured by most mathematical statisticians, and has even been carried out with non-psychological data. At the beginning of the century Pearson, who was responsible for introducing the method in anthropology, criticized all attempts to work with reduced test-variances or reduced self-correlations. Burt (1) (2) (3) has carried out various 'full' analyses, both by weighted and by unweighted summation. The method has, however, rarely been applied in recent times in the field of mental testing ; and Thomson (7) (8) (9) has once more raised the advantages of making a complete analysis of the data. The argument of the present paper closely follows his views.

With psychological data there are considerable advantages in reducing tests or traits exclusively to common factors, and making the factors as numerous as the traits : a series of inter-related tests, each of approximately equal importance, is replaced by precisely the same number of fully independent factors, a few of which account usually for the greater part of the whole variance. Although the number of variables remains unchanged, such an analysis leads to an economical ordering of the variables and a gain in conceptual clarity.

II. OBJECTIONS TO SPECIFIC FACTORS

(a) *Mathematical.* The more usual methods require a specific factor to be introduced for each test. This leads to several disadvantages, whether factors are regarded as purely mathematical components, i.e., as abstract 'factors,' or whether they are treated as psychological entities, i.e., as concrete 'abilities.'

Four objections may be raised.

(i) One of the objects of factor analysis is to achieve parsimony and economy of description by extracting fewer common factors than there are tests. But in this form the postulate of parsimony is, in one sense, an indulgence, since, while economizing in the number of common factors, we nevertheless assume in addition a specific factor for each one of the tests. Thus the total number of factors actually becomes greater than the number of tests.

(ii) Since, by including specifics, the ordinary analysis takes out more factors than tests, we are unable to make an accurate estimate of the factor scores of persons, because we now have more unknowns than equations to solve them. This means that the ordinary practice fails to give a wholly satisfactory mathematical solution to the factorial problem.

(iii) The results of factor analyses are as a rule regarded as sufficiently satisfactory if they reproduce the correlations within the limits of error. This does not seem to be an adequate criterion, since the amount of error may be considerable. By taking very small samples most correlation tables could be explained, within the limits of error, by a single general factor only. Indeed some of Spearman's early analyses could be criticized on this

ground. Analyses of this kind, in which owing to the smallness of the sample only a few factors are significant, and in which the correlations are reproduced with a wide margin of error, seem plainly unsatisfactory. We should aim rather at reducing accidental error than at allowing for it. It is true that in most methods of analysis we can approach perfect reproduction of the correlation coefficients by iteration of the calculations. But in practice most workers do not carry out many iterations because of the labour involved ; and in any case perfect reproduction of the figures given is impossible when only a limited number of common factors is used, no matter how many iterations are carried out. The important point is that, whether the factors extracted are significant or non-significant, they cannot reproduce the correlation matrix *exactly* unless they are as numerous as the tests.

(iv) The usual practice of substituting fractions for unity in the diagonal cells of the correlation matrix means that, before the analysis is begun, we have allotted a specific factor to each test and largely determined its size.

(b) *Psychological.* If factors are regarded as concrete psychological entities, i.e., as abilities rather than as mathematical components, the following points may also be urged :

(i) By including a number of sufficiently similar tests in a battery any specific factor can be turned into a group factor, or even a general factor. And, given a number of traits which are said to be governed by a general factor only, it is always possible to add one or more tests or traits which do not contain that factor, and so convert the general factor into a group factor. There is no absolute distinction between general and specific factors.

(ii) With psychological data, the common factors frequently account for no more than 50 or 60 per cent. of the total variance. As a result a test may have its highest loading in the specific factor. Thurstone (11), for example, extracts twelve primary group factors from fifty-seven tests. Of these twenty-eight have a higher loading in the specific than in any other factor ; and thirty-six of the specific loadings are greater than .50.

As has frequently been pointed out, the emphasis given to specific factors depends largely on the tests which the investigator chooses to analyse. The high specificity found in so many analyses arises from the fact that the batteries used do not include closely related tests. For purposes of vocational and educational selection, we avoid duplicating the test material. Such batteries, however, will not give a correct picture of the relative importance of specific elements in the total structure of ability. Were it feasible, an analysis of all known tests in a single battery would presumably reveal a very large number of group factors and small specifics. We have therefore to distinguish between the specific factors found in batteries assembled for practical purposes and the specific abilities in which general psychology is interested.

It is on these grounds that the factorial approach to the problem of the actual components in ability seems inadequate to the mathematician and devoid of plausibility to the general psychologist, despite the outstanding contributions that have been made by this method in the applied field.

III. ADVANTAGES OF THE FULL ANALYSIS

To be completely satisfactory from a mathematical standpoint, the solution should have as many significant factors as there are tests. This means that it will be necessary first, to carry out a full analysis, and second, to work with a sample large enough to ensure the significance of each factor. This would have the following advantages :

- (i) There would be maximum economy in the total number of factors ;
- (ii) A person's factor scores could be calculated with precision instead of being estimated approximately ;
- (iii) The inter-test correlations could be reproduced exactly ;
- (iv) The specific factors of the ordinary analysis would in general be replaced by near-specific factors, in which a large proportion of the factor variance is concentrated in a single loading. The idea of a 'near-specific factor,' which is largely confined to the performance of a single test, but not completely devoid of loadings in the other tests, seems as reasonable a psychological concept as the completely specific factor with which we are more familiar.

IV. DISADVANTAGES OF THE FULL ANALYSIS

The arguments against a full analysis may perhaps be summarized as follows.

(i) Experience has shown that for most practical purposes, there is little point in extracting as many factors as there are tests merely for the sake of completeness. In applied psychology, the object of the analysis is usually better served by restricting the number of common factors. And the suggestion that the size of the sample should be increased until all the factors are significant is a counsel of perfection since the number of subjects required to make the later (and less important) factors significant might be enormous.

(ii) With psychological measurements some of the specific variance is due to unreliability or to the error variance peculiar to a given test applied on a given occasion, e.g., the effects of the temperature of the room, the vigilance of the supervisor, street noises, or early morning freshness—conditions which may disturb one test but not another. Yet the full analysis insists on expressing all this variance in terms of common factors.

(iii) It is reasonable to suppose that, with the batteries ordinarily used, specifics unique to the particular tests on a particular occasion do in fact exist. Even in a highly homogeneous set of tests there must sometimes remain specific tricks and knacks, useful for one test only, much in the same way as skill in handling a particular padlock or bicycle does not help us with other forms of the same apparatus. It can scarcely be supposed that no unique element enters into the ability to perform any test whatsoever. Hence, if the ordinary simple summation analysis appears to over-emphasise the importance of specific abilities, the full analysis is also at fault in eliminating all specifics on all occasions.

Nevertheless, it must, I think, be admitted that the over-all picture of ability is less distorted by the full analysis. Although such an analysis is too laborious to be recommended as a routine method in applied psychology, it is of considerable theoretical interest to the general psychologist. As we have seen, it provides the only adequate mathematical solution to the factorial problem, and at the same time gives a more reasonable, if not perhaps completely satisfactory, account of the specific elements in the over-all structure of ability.

V. THE ENQUIRY

The computational labour of the full simple summation analysis as usually carried out can be avoided by the triangular method (3) (5) (10). The purpose of the present investigation is to make a closer study of the type of solution to be expected if this method is applied to data in the field of ability. The results of a full triangular analysis of a typical battery are compared with results already obtained for the same battery by Vernon (12) with the simple summation and group factor methods. The tests were (1) Matrix, (2) Abstractions, (3) Dictation, (4) Arithmetic and Mathematics, (5) Bennett Mechanical Information, (6) Electrical and Mechanical Information, (7) Squares, (8) Memory for Designs. The subjects were 500 seamen recruits.

This battery was chosen because (i) it included a wide range of the type of tests commonly used in factorial studies in educational and vocational psychology, viz. verbal and non-verbal intelligence, scholastic ability, scientific information and spatial discrimination; (ii) Vernon (12) had already carried out several analyses of this battery; (iii) the battery was known to be complex factorially; (iv) it was applied to a fairly large sample; (v) the amount of common factor variance was fairly typical of psychological data.

VI. THE TRIANGULAR METHOD OF ANALYSIS

As Burt has pointed out, a triangular analysis is easier to carry out than a simple summation analysis (3) (5) (10). It makes a convenient starting point for rotation, since nearly half the loadings are necessarily zero. But, although it yields a perfect reproduction of the original correlation matrix, it has serious theoretical disadvantages owing to the arbitrary

nature of the loadings. The factor matrix has a standard pattern consisting of a triangular block of zero entries (Table I), with no zeros in the first factor, 1 zero in the second, 2 in the third and $n - 1$ in the n th factor, this last being a specific. Mathematically the choice of the order in which the factors are extracted is entirely free. In the present enquiry, the eight tests were written in order of their average correlation with the other seven tests. This gives as much weight as possible to the variance of the earlier factors, and has the effect of increasing the number of small loadings in the later factors.

TABLE I. TRIANGULAR ANALYSIS

Factor	I	II	III	IV	V	VI	VII	VIII
Abstractions	1.000	.000	.000	.000	.000	.000	.000	.000
Arithmetic	.616	.788	.000	.000	.000	.000	.000	.000
Mech. Inform. ..	.463	.135	.876	.000	.000	.000	.000	.000
Matrix	.540	.181	.126	.812	.000	.000	.000	.000
Memory for Designs ..	.400	.193	.295	.154	.832	.000	.000	.000
Squares	.392	.114	.259	.219	.285	.798	.000	.000
Elec. Inform.	.393	.145	.452	(-.063)	(.089)	(-.028)	.780	.000
Dictation	.570	.339	(-.030)	(.031)	(-.053)	(-.050)	(.031)	.742
Sum of squares ..	2.678	.858	1.142	.736	.784	.641	.609	.551
							Total :	7.999
Percentage contribution to variance	33.5	10.7	14.3	9.2	9.8	8.0	7.6	6.9
							Total :	100.0

Note.—Loadings printed within parentheses are non-significant.

The factors are not as difficult to interpret as might be expected. Some are sufficiently comprehensive to give useful information about the battery of tests as a whole, while others resemble specifics. The first factor is a general factor of the usual type; the second is scholastic, the third is (largely) scientific information, factors IV–VII are near specifics and the last factor is purely specific.

The triangular matrix was rotated to obtain a maximum number of (i) positive and (ii) near zero loadings, lying between $+ .100$ and $- .100$, i.e., the criterion chosen was similar to, but less exacting than, simple structure. The computer was unaware of the identity of the tests. The figures obtained are shown in Table II.

VII. RESULTS

Vernon (12) carried out simple summation and group factor analyses of the data, with the results shown in Table III. (The writer is much indebted both to him and to the Senior Psychologist's Department of the Admiralty for permission to reproduce these unpublished tables.)

Both analyses yield a general factor and classify the tests into four groups presumably corresponding to (i) spatial ability (Squares, Memory for Designs), (ii) scientific information (Mechanical Information, Electrical Information), (iii) scholastic ability (Dictation, Arithmetic, Abstractions), and (iv) non-verbal intelligence (Matrix). The main features of the rotated triangular solution are first of all that, despite its limitations, it clearly emphasizes the four factors found by Vernon. The first four factors closely resemble those obtained by Vernon with a group factor analysis. The chief discrepancy is that in his second factor Abstractions is added to the Dictation and Arithmetic tests, thus making a verbal-educational factor. Secondly, the specific factors of the ordinary analysis are replaced by near-specific factors. Simple structure could not be attained owing to the prominence of the basic factor.

TABLE II. TRIANGULAR ANALYSIS : ROTATED LOADINGS

Factor	I	II	III	IV	V	VI	VII	VIII
Arithmetic	.688	.714	(-.078)	(.009)	(.071)	(-.055)	(.005)	.000
Dictation	.586	.281	(-.076)	(-.083)	(.066)	(-.098)	(.026)	.742
Mech. Inform.	.584	(.088)	.798	(.100)	(.008)	(.087)	(.015)	.000
Elec. Inform.	.459	(.098)	.383	(.064)	(-.061)	(-.005)	.794	.000
Memory for Designs ..	.513	(.061)	(.100)	.844	(.016)	(.016)	(.088)	.000
Squares	.501	(.038)	(.028)	.287	(.100)	.808	(.032)	.000
Matrix	.575	(.043)	(.032)	(.065)	.811	(.061)	(.010)	.000
Abstractions	.981	(-.100)	(-.100)	(-.098)	(-.010)	(-.069)	(-.005)	.000
Sum of squares ..	3.175	.624	.818	.829	.687	.683	.640	.551
							Total :	8.007
Percentage contribution to variance	39.7	7.8	10.2	10.4	8.5	8.5	8.0	6.9
							Total :	100.0

TABLE III. SIMPLE SUMMATION AND GROUP FACTOR ANALYSES

Tests	A. Simple Summation			B. Group Factors			
	I	II	III	I	II	III	IV
Squares	.605	-.243	-.407	.54	—	—	.74
Memory for Designs ..	.634	-.241	-.205	.63	—	—	.22
Mech. Inform.	.681	-.298	+.139	.65	—	.50	—
Elec. Inform.	.616	-.362	+.474	.53	—	.50	—
Dictation	.606	+.460	+.154	.46	.70	—	—
Arithmetic	.723	+.314	+.044	.65	.45	—	—
Abstractions	.764	+.253	+.009	.72	.33	—	—
Matrix	.635	+.112	-.198	.68	—	—	-.07

VIII. SUMMARY

1. It is argued that, although factor analysis has made an outstanding contribution to applied psychology, the factorial account of the specific elements in ability seems neither adequate to the mathematician nor plausible to the general psychologist. A full analysis in which as many factors are extracted as there are tests, gives a more reasonable, although not a completely satisfactory, account of the importance of specific elements in the total structure of ability.

2. The results of a full triangular analysis of a typical battery of tests of ability are compared with the results obtained for the same battery by the ordinary simple summation and group factor methods. After graphical rotation to obtain a maximum number of positive loadings and a maximum number of small loadings, the final set of loadings (a) emphasizes the same factors as those previously indicated by other procedures ; (b) gives a ' basic-plus-group ' factor solution ; (c) substitutes near-specific factors for the ordinary specifics.

REFERENCES

1. Burt, C. (1947). 'A Comparison of Factor Analysis and Analysis of Variance.' *Brit. J. Psychol., Stat. Sect.*, I, 3.
2. Burt, C. (1949). 'Alternative Methods of Factor Analysis.' *Brit. J. Psychol., Stat. Sect.*, II, 98.
3. Burt, C. (1951). 'The Numerical Solution of Linear Equations.' *Brit. J. Psychol., Stat. Sect.*, IV, 31.
4. Burt, C., and Banks, C. (1947). 'A Factor Analysis of Body Measurements for British Adult Males.' *Ann. Eugen.*, XIII, 238.
5. Frazer, R. A., Duncan, W. S., and Collar, A. R. (1947). *Elementary Matrices*. Cambridge : University Press.
6. Hartog, P. S., and Rhodes, E. C. (1936). *Marks of Examiners*. London : Macmillan.
7. Thomson, G. (1948). *The Factorial Analysis of Ability*. London : University of London Press.
8. Thomson, G. (1949). 'The Nature of the Mind's Factors,' *The Advancement of Science*, VI, 267.
9. Thomson, G. (1950). 'Note on Sir Cyril Burt's Paper on Differential Weighting,' *Brit. J. Psychol., Stat. Sect.* III, 126.
10. Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago : University Press.
11. Thurstone, L. L. (1937). *Primary Mental Abilities*. Chicago : University Press.
12. Vernon, P. E. (1945). Notes on Statistical Method. (Unpublished.)

THE REDUCED CORRELATION MATRIX

A Reply to Dr. Warburton's Criticisms

By CHARLOTTE BANKS and CYRIL BURT

University College, London

I. *Problem.* II. *The Principle of Economy.* III. *The Status of Specific Factors.* IV. *Mathematical Objections to Reduced Correlation Tables.* V. *Correction for Specificity.* VI. *The Substitution of Near Specifics.* VII. *Summary and Conclusions.*

I. PROBLEM

The practice of substituting reduced self-correlations for self-correlations of unity has recently been questioned by several writers (e.g., Thomson, 1949, Kendall *et al.*, 1950); but the reasons for or against the procedure have never been systematically examined. Dr. Warburton has summarized the objections most frequently raised, and decides in favour of a return to the 'full test-variances,' i.e., to self-correlations of unity, such as were adopted by Karl Pearson. His arguments are based partly on abstract or general considerations, both psychological and mathematical, and partly on a concrete analysis of what he believes to be a 'typical battery.' The illustration that he uses consists of correlations for eight selection tests applied to 500 'seamen recruits' (Table I). The tests themselves, or at least the majority of them, are much the same as those included in the larger batteries of tests applied to 1,000 Army recruits, and analysed by one of us (Banks, 1949) in an earlier issue of this *Journal*. Similar factorizations were also made by our colleagues for analogous test-data obtained from Naval recruits. All these early analyses appeared to indicate that the use of reduced self-correlations furnished the most economical and consistent explanations. The method itself was originally put forward by Burt (1909, pp. 94, 177; 1915, p. 696; 1917, pp. 53f.), who adopted it from the start for most of his investigations. Accordingly, it seems incumbent on us to defend in fuller detail the methods that we have used.

First, however, it should be expressly stated that we do not maintain (as Thurstone and others have done) that reduced communalities should be automatically adopted on *all* occasions. Both of us have already published results obtained with a 'full factor analysis' (e.g., Burt, 1949b, Banks, 1954). The problem, as we see it, is not to decide which method is the best for every conceivable purpose, but rather to determine under what conditions the one method or the other should be employed. Still less do we wish to deny all cogency to Dr. Warburton's arguments. His conclusions would be altogether reasonable if the assumptions on which they rest could be accepted; and, as he points out, his assumptions are tacitly if not explicitly adopted by most factorists. Nevertheless, we believe that, as commonly stated, they imply several misconceptions. Thus the problems emerging from Dr. Warburton's discussions are really twofold: (i) what precisely are the assumptions that have led to the use of reduced self-correlations in certain cases and not in others? (ii) how far can those assumptions be justified (*a*) on logical and (*b*) on psychological grounds?

II. THE PRINCIPLE OF ECONOMY

As in other problems of factorial procedure, the differences in the methods advocated spring largely from differences over scientific methodology; and these raise basic issues which have hitherto been somewhat neglected (see this *Journal*, I, p. 2). Dr. Warburton begins by reminding us that "one of the objects of factor analysis is to achieve parsimony of description"; and this he takes to mean that the number of factors extracted should be less than the number of tests. A similar interpretation of the 'principle of parsimony,' as

Thomson has also pointed out (1949, p. 269), is apparently accepted by most other psychologists, including Thurstone (1947).¹ Yet it does not appear to be wholly correct. Why should the number of *tests* be singled out as the governing condition? And why should the figure to be minimized be the number of *factors* rather than the number of numerical constants?

Suppose we have three sets of measurements giving the positions of a hundred stationary balloons: three common factors would be necessary to describe the results. Now add three more sets of measurements taken from different points of observation: three factors might still suffice. Similar cases can readily be imagined with chemical or physiological tests. Instances of this kind strongly suggest that often it is the nature of the tests or of the sets of measurements rather than their actual number that determines how many factors are to be extracted. In psychology, for example, the number of group factors required to explain a given correlation matrix is always more than the number of bipolar factors. But no-one would argue that the group factor hypothesis should be rejected simply for that reason. Our choice would rather depend on the nature of the traits: most factorists prefer group factors for cognitive traits, but are ready to accept bipolar factors for temperamental traits.

The principle popularly known as Ockham's razor² may perhaps best be formulated in general terms as follows: "In attempting to explain a given set of empirical phenomena, it is a breach of scientific method to introduce more assumptions than are necessary." What, then, are the implications of this principle as applied to data such as Dr. Warburton has in mind?

Let us begin by supposing, as he does, that what we want to 'describe' (i.e., to fit with appropriate figures deduced from our hypothesis) are, not the observed test-measurements, but the observed correlations. These observed correlations, we may presume, are, with

¹ Thomson repeatedly refers to the 'principle of parsimony' in the course of his paper. But it is not quite clear how he himself interprets it. His own point, however, is rather that the factorists who appeal to it are not altogether clear about what they understand by it.

² Previous factorists, in discussing this maxim, have spoken of it as "a principle introduced by modern scientists"; others describe it as "a principle first introduced by Occam." Neither statement is correct. The *lex parsimoniae* was introduced, not by "modern scientists," but by scholastic scientists; and it is much older than William of Ockham. It is a methodological principle distinctive of the English empiricists, and notably of the Franciscan school that flourished at Oxford during the 13th and 14th centuries—a school consisting largely of experimentalists who were constantly at loggerheads with the Dominican school that dominated continental philosophy and claimed St. Thomas Aquinas as its most eminent exponent. The earliest explicit formulation would seem to be that of Robert Grosseteste: "Melior est demonstratio, aliis circumstantibus paribus existentibus, quae eget paucioribus suppositionibus ex quibus demonstratur" (Commentary on Aristotle's newly discovered *Posterior Analytics*, i, 17, f.17^v, a passage which, as Grosseteste points out, really implies some such principle). Grosseteste was born just a hundred years after the battle of Hastings; became the first Chancellor of the University; and may be regarded as the founder of the experimental school at Oxford. Roger Bacon (d. 1249), who nowadays is perhaps the best known member of this school, repeatedly cites the principle, more or less in Grosseteste's words. The version that I cited in *Factors of the Mind*, "Frustra fit per plura quod potest fieri per pauciora," was first enunciated by Duns Scotus (a member of the same circle, d. 1308), and became a favourite maxim with his followers, the 'dunces.' Ockham (who was born at Ockham in Surrey about 1290 and died during the Black Death) was a pupil of Duns, and, politically as well as philosophically, one of the most influential of the schoolmen. He uses Duns's version of the principle, and also gives a shorter one of his own: "Pluralitas non est ponenda sine necessitate" (*Quotlibeta*, v, 5). But the form most frequently cited—e.g., by H. Jeffreys, who wrongly attributes it to Ockham (*loc. cit.*, p. 224), and by other writers on mathematical physics—is "entia non sunt multiplicanda praeter necessitatem." This, however, is not found until the 17th century, and was apparently suggested by John Ponce of Cork, a Scotist. (The interested reader will find an entertaining discussion of the whole subject in W. M. Thornburn, 'The Myth of Occam's Razor,' *Mind*, N.S. XXVII, 1918, pp. 345 sqq.; his account, however, requires some correction in the light of more recent historical research on the work of the early Oxford philosophers.)

To avoid possible misconception, may we add that we should prefer to treat the 'law of parsimony,' not as a basic postulate, but as a corollary to the postulate of maximum likelihood? To put the argument briefly and therefore rather crudely, it may be said that a hypothesis containing fewer assumptions must be 'better' than one that involves more, because, if the *a priori* probability of a single assumption is x , then the probability of n assumptions will be x^n , and since x is a fraction, x^n must diminish with the size of n .

few exceptions, positive and significant : else we should hardly trouble to factorize them. In such a case we shall plainly be justified in assuming at least one general factor. With this provisional assumption we can easily deduce a set of hypothetical correlations which will fit (more or less accurately) the correlations actually observed. The question now arises : how far do these hypothetical values deviate from the observed coefficients, and will the deviations or 'residuals' justify us in postulating one or more additional factors? The principle of economy, as we have stated it, warns us not to assume a second factor if the observed correlations can be adequately explained by assuming a single factor only. It follows that we must make this first factor fit the observed coefficients as closely as possible. In other words, we must take, not any common factor, but what has sometimes been termed 'the highest common factor' (Burt, 1909). The obvious criterion for closeness of fit will be the numerical sum of the deviations, or better still the sum of the squares of the deviations. Having minimized this square-sum, we must next test the deviations for statistical significance. The test proposed was Pearson's chi-squared criterion, which also turns on the square-sum. Only if the deviations prove to be statistically significant shall we be warranted in introducing a further factor. If they are, we shall then apply the same procedure to determine whether a third factor is warranted ; and so on.

This explanation should make the difference between Thurstone's interpretations and Burt's perfectly clear—a difference which has apparently been overlooked in Dr. Warburton's comparisons. Thurstone seeks the smallest number of common factors that would fit the side correlations, i.e., a number identical with the rank. Burt seeks the smallest number of common factors that are statistically significant, regardless of the rank. As a rule, the number of factors required by the latter method is less than that required by Thurstone's. Thomson, referring to Thurstone's centroid formula, notes that "this method of finding saturations had been devised and employed many years previously by Burt, though it is not quite clear how he would fill in the diagonal cells" (1948, p. 25). The working procedure originally used was to take the reliability coefficients and to reduce them by preliminary trials until they were only slightly in excess of the adjacent coefficients when the table was arranged as nearly as possible in hierarchical order : after 'correcting for specificity,' the residuals were then tested by the probable errors (Burt, 1917). Later, when larger samples of the population were assessed, more rigorous methods were put forward, all depending on successive approximation. Thurstone, it will be remembered, prefers to insert the largest coefficient in each column—a simple mechanical rule which appeals to the beginner, but proves usually to over-estimate the number of significant factors. Here, however, it is not the choice of procedure but the choice of principle that is primarily at issue.

III. THE STATUS OF SPECIFIC FACTORS

Dr. Warburton's second objection turns on the fact that, with reduced self-correlations, we postulate, not only r common factors ($r < n$), but also n specifics. Hence the total number of factors is really $n + r$; and this is plainly greater than the number of tests. To overcome this objection he proposes to substitute for the n specifics a smaller number of what he calls 'near specifics.' A 'near specific' is a common factor which has most of its variance (but not all) concentrated in a single test. This, he thinks, is "as reasonable a psychological concept as the completely specific factor."

Now in dealing with problems of statistics it is scarcely correct to treat the different kinds of factor as different kinds of 'psychological concept.' The factors with which the computer is concerned are not (as Dr. Warburton assumes) 'concrete psychological entities.' Indeed, they are not 'concrete entities' at all : they are logical or methodological concepts rather than psychological.

There is so much confusion over this point that it is desirable first of all to ask what precisely we mean by the term 'specific factor.' The earliest classification of factors corresponded to the logical classification¹ of predicative propositions : (i) if we can say "all

¹ This is virtually Aristotle's classification (*De Interpretatione*, 7 ; but cf. *An. Pri.*, 24^a 17–22 and 43^a 25–32). The traditional terminology will be found in any text-book of logic, e.g., Mill, I, iv, § 4 : 'Of Propositions : Universal, Particular, and Singular.' These are simply the Latin terms, fixed apparently by the authority of the text-book of Petrus Hispanus (afterwards Pope John XXI), *Summulae Logicales* (c. 1260). The earliest Latin commentators vary somewhat in their translations of the Aristotelian terminology, often using *generalis* for *universalis* and *unicus* for *singularis*.

the traits we have tested possess this factor," then the factor is 'universal' or 'general'; (ii) if we can only say, "some of the traits possess the factor," then it is 'particular' or 'special'; (iii) if the most we can say is "this one trait possesses the factor," then it is 'singular' or 'unique.' Moreover, if each trait has been tested on one occasion only, then we have also to take into account (iv) the 'unreliability' of a single testing, i.e., the 'chance' or 'error' factors.¹ When (as is now very usual) all the tests are applied at the same sitting, the errors of different tests are likely to be correlated. However, if due precautions are taken, these correlations will usually be relatively small; and most recent writers treat them as non-existent or at least negligible. In that case, the factors of unreliability assumed in our statistical model will also be unique. Dr. Warburton would apparently agree, since he includes them in what he terms 'specificity.'

Let us call this earlier classification and this older terminology the 'British.' A somewhat similar scheme—reached, it would seem, not from logical but rather from psychological considerations—is implied by the nomenclature used by Holzinger (1941) and Thurstone (1947): indeed, Thurstone himself frankly identifies 'factors' with 'faculties' (*loc. cit.*, p. 70). Let us call this scheme with its terminology the 'American.' The analogies between the two are shown in the table below.

TABLE I. THE SUBDIVISION OF TEST VARIANCE ACCORDING TO THE FOUR FACTOR THEORY

<i>British Nomenclature</i>	<i>American Nomenclature</i>	<i>Variance</i>	<i>Symbols Brit. Amer.</i>
A. Common	A. Common	A. Communality	σ_c^2 , h_j^2
1. Universal or General	1. General	1. General Factor	σ_g^2 , g_j^2 or a_j^2
2. Particular or Special (Bipolar or Group)	2. Other Common Factors	2. Variance of Supple- mentary Factors	σ_b^2 , b_j^2
B. Individual	B. Unique	B. Uniqueness	σ_u^2 , u_j^2
3. Singular or Unique	3. Specific	3. Specificity	$\sigma_{s_j}^2$, s_j^2
4. Error or Unreliability	4. Error	4. Unreliability	$\sigma_{e_j}^2$, e_j^2

The analysis of the total variance of a test j into the component factor variances may thus be represented by the following equations (Hartog and Burt, 1936, p. 259, eqns. v *et seq.*).

$$\sigma_j^2 = \sigma_g^2 + \Sigma \sigma_b^2 + \sigma_u^2 + \sigma_{e_j}^2.$$

If the tests are in standard measure, as the American notation assumes, we have

$$\begin{aligned}\sigma_j^2 = 1 &= g_j^2 + \Sigma a_j^2 + s_j^2 + e_j^2, \\ &= h_j^2 + u_j^2.\end{aligned}$$

Thus the 'reliability coefficient' of a test will be $r_{jj} = 1 - e_j^2$, and the 'reduced self-correlation' $\bar{r}_{jj} = 1 - s_j^2 - e_j^2$.

¹ This method of classifying factors was put forward by Burt in several earlier papers: the fullest discussion is to be found in his Memorandum to the International Institute Examinations Inquiry reprinted in Hartog *et al.* (1936).

When the group of traits represented a class or *genus* (e.g., all cognitive or all emotional), the factor was called 'generic' or 'general': when they represented a subclass or *species*, it was called 'specific.' These were the meanings attached to the terms by 19th century psychologists (cf. Bain, *Senses and Intellect*, pp. 639f., 'Classifications of Intellectual Powers'). Spearman, however, maintained that the only non-general factors were 'specific' to single tests. Hence the word got transferred to the third type of factor; and those who still believed in a third type of factor had to substitute some phrase like 'special' or 'group' factor, to avoid ambiguity. In articles for teachers and others the academic language of the logician was usually dropped; and during recent years almost every writer on factor analysis has adopted a nomenclature of his own.

In psychological work, no test can be treated as entirely free from 'unreliability' and 'specificity' (in the sense defined by the table). The presence of 'unreliability' was clearly demonstrated in early researches where it was customary to apply every test at least twice: with samples of reasonable size, the correlation between the two applications (the 'reliability coefficient') was always significantly less than unity. The presence of 'specificity' was demonstrated by entering the two tests separately in the correlation table: a significant group factor almost invariably appeared. When the measurements for each pair of tests were pooled, so that each type of test was entered only once in the correlation table, then the group factors were converted into 'singular' factors, i.e., 'specific factors' in the narrower sense. In more recent factorial researches, where it is quite exceptional for tests to be applied twice, it is impossible to decompose the 'uniqueness' into 'specificity' and 'unreliability': for every test the two are merged into a single factor. Consequently, the phrase 'specific factor' has come to be used in a somewhat wider sense, and 'specificity' has been treated as synonymous with 'uniqueness.' Apart from one or two minor inconsistencies, this is the usage that Dr. Warburton has adopted. Hence, in discussing his arguments, we shall keep to his terminology; and the context will make the meaning perfectly clear (cf. also Thomson, 1949, p. 269).

If this interpretation is correct, it becomes obvious that what Dr. Warburton calls 'specific factors,' unlike the other factors, are not *postulated* entities, with which we can dispense if we wish. We accept their presence, not on grounds of an *ad hoc* hypothesis, which our enquiry is designed to check, but on grounds of antecedent knowledge. When applying the principle of parsimony, therefore, it would plainly be wrong to count them among our hypothetical concepts. Thus, with the eight tests in Dr. Warburton's example, we surely know in advance, without further observational or statistical check, that each test must contain a 'unique' or 'specific' factor not shared by any other test. If in the end we are able to explain these eight sets of test-measurements in terms of only three significant common factors, then our hypothetical model offers a greatly simplified description of the individual measurements; and the factors so established are, in accordance with sound methodology, 'overdetermined.'

Dr. Warburton goes on to contrast his own 'full factor analysis' with a summational analysis and a group factor analysis obtained for the same data by Professor Vernon. Although both have eight specific factors, the summational method yields only three common factors (one general and two bipolar), whereas the group factor method yields four (one general and three group). Most psychologists would almost certainly prefer the group factor analysis. Yet the law of parsimony, as interpreted by Dr. Warburton, requires us to prefer the former.

But here we must question Dr. Warburton's interpretation of the principle. Surely what the 'law of economy' requires us to minimize is not the number of hypothetical factors but the number of hypothetical constants.¹ We therefore suggest the following formulation of the principle when applied to factor analysis: "other things being equal, that factorial hypothesis is to be preferred which postulates the smallest number of arbitrary constants."² Let us apply this to his table.

As we have already seen, the general type of factorial model that we put forward for verification will usually be determined from the very outset by our antecedent knowledge of the field with which we are dealing and by the kind of questions we wish to put. If these had suggested that the most suitable classificatory scheme for Dr. Warburton's tests was

¹ We believe that this argument would largely meet Professor Thomson's criticisms. He is "sceptical about the procedure of making the specific factors *as large as possible* by an automatic mathematical device . . . without any psychological consideration whatever being given to the question" (the reference is apparently to Thurstone's procedure for securing the minimum rank permitted by the correlations actually observed). Nevertheless Professor Thomson would also "like to see existing tables of correlations re-analysed with full variances."

² This would seem to be in keeping with the interpretation accepted in other branches of science. Physicists, for example, measure the 'complexity' of a differential equation by taking the sum of the order, degree, and number of coefficients; and, other things being equal, prefer the least complex equation. 'Order' applies only to differential equations; and the equations of ordinary factor analysis are all of the first degree. Hence here 'complexity' depends solely on the number of coefficients.

one that involved general and bipolar factors, then, since the number of persons tested was only 500, we might expect three summational factors to be significant but no more. Such a hypothesis would involve $8 + 7 + 6 = 21$ constants, namely, the factor saturations. However, our antecedent knowledge makes it at least equally legitimate to adopt positive group factors; and, on attempting a group factor analysis, we discover that as good a fit can be obtained with 15 saturations only. After all, therefore, the results of the group method seem preferable. And this conclusion is confirmed by the earlier analysis carried out with a larger battery of tests and with a larger number of persons: the bipolar matrix contained in all four factors and 46 saturations; the group factor matrix (which consisted of subdivided factors) contained as many as seven factors but only 34 saturations (Banks, 1949).

IV. MATHEMATICAL OBJECTIONS TO REDUCED CORRELATION TABLES

There is, however, a third objection, on which Dr. Warburton lays greatest stress. Any solution which entails fewer factors than tests would, he contends, be repugnant to the mathematician. He is presumably referring to the views expressed by Professor G. A. Barnard and others at a recent symposium on factor analysis arranged by the Royal Statistical Society (1950). Professor Kendall in his opening paper began by observing that the method of principal components, which, at least in its original form, involves a 'full' factor analysis, would be "the approach which the statistician would first consider," and "the one to which the psychologist might be expected to give pride of place." He himself does not, however, reject the introduction of specific factors or the use of self-correlations smaller than unity: he merely points out that certain "auxiliary hypotheses" are needed to make the solution determinate.

Barnard's criticisms are more drastic. He declares that he has "not met any problem dealt with by factor methods which could not have been better dealt with by classical multivariate methods" (among which he includes the original method of principal components since that, he tells us, is simply "a natural extension to vectors of the idea of multiple correlation"). He adds that his criticisms are meant primarily to apply to "most of the factor work received from the other side of the Atlantic"; and he excludes the work of certain British factorists. However, the views of the various speakers were by no means unanimous; and one of us (Banks, 1954) has since re-analysed Professor Kendall's main tables of correlations, using reduced self-correlations and supplementary factors as well as a general factor, and has shown that this procedure gives if anything a clearer interpretation of the data.

Both methods, however, undoubtedly have their merits and their uses; and before a choice is made it becomes essential to ask what, in any particular enquiry, is the computer's purpose. If the summational analysis is needed merely to throw the data into a more convenient form for subsequent manipulation, then we may readily agree that a 'full analysis' will often be justifiable. But that is not the usual aim of factor analysis as applied by the psychologist. The preliminary summational analysis—whether carried out by the more rigorous method of weighted summation (used in seeking principal components) or by the speedier method of simple summation (used as a convenient substitute for the more laborious procedure)—has for its main objects (a) to determine, with the aid of a factor pattern of positive and negative signs, how the tests are to be classified, (b) to discover the minimum number of significant factors or significant factor saturations, and (c) to remove the contributions of error or specificity, such as are incorporated in the unique factors.

But, according to Dr. Warburton, the mathematician is not interested in removing certain constituents or in 'allowing for error': his sole concern is that the factor matrix should yield a "perfect reproduction of the correlation coefficients." To do this, says Dr. Warburton, "will be impossible when only a limited number of common factors is used." Here he seems to be in error: a perfect reproduction of the figures in the correlation matrix can always be achieved with fewer factors than tests, provided appropriate self-correlations

are inserted.¹ However, though it is certainly possible to reproduce a given set of correlations, it will not be possible to reproduce the initial data, i.e., the original test-measurements. And if our primary object is to find a set of uncorrelated factor-measurements which will be an exact equivalent² of the correlated test-measurements, then we must use the same number of factors as there are tests.

One difficulty in Dr. Warburton's proposal seems obvious at the outset. The numerical values with which the factorist is concerned—his observed test-measurements and the correlations derived from them—are not pure numbers, but empirical measurements; and it is a truism that any observed measurement must involve an inevitable error resulting from the essential limitations of every measuring instrument. Measure the diagonal of a square with unit sides or the circumference of a circle of unit radius: the measurements obtained will never be, and indeed never can be, exactly equal to $\sqrt{2}$ or to π . No matter how fine we make our scale, we can read it only to the nearest multiple of a certain constant called its 'step.' Thus, if we take the step as unit, and suppose the observed value to be n , the true value will be, not n , but $n + x$, where n is a whole number and x is an unknown quantity lying between $\pm \frac{1}{2}$. When Dr. Warburton tells us that the correlation of Dictation with Arithmetic is .618, that only means that it lies somewhere between .6175 and .6185. By carrying the calculation to further decimal places, we might reduce the amount of inaccuracy, but we can never abolish it altogether. Accordingly, since we are bound to 'allow for' this kind of error, there is no reason why we should not allow for other kinds of error; and, when the measurements themselves are of necessity imperfect, there can be little virtue in aiming at

¹ Dr. Warburton writes as though the number of factors needed to reduce the correlation table to its minimum rank might differ quite arbitrarily with different tables of the same size. Such differences do in fact occur with the working methods adopted by many computers, e.g., by those who insert in the leading diagonal the largest figures in each column. But these approximate procedures do not really yield the minimum rank.

In theory the number of factors required can be deduced as follows. The symmetric matrix R will have a rank of r , if (a) at least one r -rowed principal minor is not zero, and (b) all the principal minors obtained from this minor by adding one row and the corresponding column are zero (cf. M. Bôcher, *Higher Algebra*, 1931, p. 57, *Theorem 1*). It is easy to see that the number of different bordered minors of this type will be $\frac{1}{2}d(d+1)$, where $d = n - r$. Now in order to transform a given table of observed correlations to a Gramian matrix of rank r , we propose to insert in the principal diagonal whatever values are suited for that purpose. This means that we have n values to determine for the self-correlations. Suppose first that the number of tests is $n = \frac{1}{2}d(d+1)$, where d is any number we like to name. Then, by the theorem just cited, we have $\frac{1}{2}d(d+1)$ independent equations with which to determine the $\frac{1}{2}d(d+1)$ unknown self-correlations (for a proof that the equations are in fact independent, see Ledermann, 1937). Under these conditions a finite number of solutions is possible; and, with the working procedures commonly used, one solution only will in general be obtained. The number of factors will then be $r = n - d = \frac{1}{2}d(d-1)$. In other words, if the number of tests is one of the triangular numbers, 3, 5, 10, 15, . . . , the number of factors will be the next smallest triangular number, 1, 3, 5, 10, With summational methods the number of independent saturations in the factor matrix will be n for the first factor, $n-1$ for the second, and so on—namely, $\frac{1}{2}n(n-1)$ in all. Where the number of tests is not a triangular number, the minimum number of factors will be the smallest integer larger than $\frac{1}{2}\{(2n+1) - \sqrt{8n+1}\}$, and the solution will not be unique. Thus with four tests we can insert in any one of the diagonal cells whatever fraction we please (so long as it is consistent with the Gramian properties of the correlation matrix), and then fix one of the eight factor saturations at zero; the remaining seven saturations will in this way be uniquely determined.

Much of this is already familiar. What has not been sufficiently noted is the wide variety of alternative solutions that are available when the number of tests is not a triangular number. As soon as they find that their calculated self-correlations agree with their assumed self-correlations, most computers ignore the possibility of other solutions. In these and other cases, where several sets of saturations will fit the observed correlations equally well, we suggest a supplementary principle, namely, that in the absence of antecedent knowledge to the contrary the specific factor variances should be assumed to be as nearly equal as the data permit.

² Dr. Warburton does not make it quite clear that the mathematician is thinking of theoretical 'perfection,' not of the actual values that would be obtained by practical calculation. From the standpoint of pure mathematics the problem is simply to find a determinate algebraic solution for a set of n linear equations in n unknowns. E.g., given $M = FP$ and $PP' = I$, we require to find $P = F^{-1}M$. This implies that F must be non-singular and therefore square. And if it is square, it must contain the same number of columns (factors) as rows (tests).

'perfect' reproduction. After all, what we really want is not 'perfect' reproduction—that would not be desirable even were it possible—but only the most satisfactory fit obtainable¹ within the limits of estimated error.

Much the same remarks apply to Dr. Warburton's claim that with his procedure "a person's factor scores could be calculated with precision instead of being estimated approximately." Thomson also points out that Hotelling's principal components "have one advantage over other kinds of factors: *they can be calculated exactly*" (1948, p. 78, *his italics*). This might be a merit in calculations required for purely mathematical purposes; but it is not a merit in statistical calculations, since there (as every text-book reminds us) one of the chief problems is to find acceptable estimates for quantities which can never be known with exactitude. A factor score by its very nature is a purely hypothetical assessment for a purely hypothetical tendency. It is bound to be approximate, and cannot possibly be "calculated with precision." Further, regarded as estimators, the expressions derived from the principal-axes technique with unit self-correlations yield, as Frisch (1934) has shown, biased estimates.²

V. CORRECTION FOR SPECIFICITY

However, Pearson's method of principal components, to which Dr. Warburton refers, began by assuming that the object is to fit, not the observed correlations, but the observed test-measurements. At first sight this would seem to provide a more plausible starting point; and it is largely because the psychological factorist has departed from this formulation of the problem that so many critics have objected to the alternative methods that are now in regular use.

Their objections can readily be met by employing a more complicated procedure. Pearson's method was based on the principle of least squares; but it was developed for dealing with physical measurements, where the element of unreliability is almost negligible. With mental measurements the element of unreliability is so great that it cannot be ignored. Now in other fields of work, when the principle of least squares is adopted, there is a recognized device for 'correcting' for the unreliability. If we borrow it for correlational work, it yields the formulæ suggested by one of us as a means of 'correcting for specificity' (Burt, 1917, 1949a). Each variable is first weighted according to its relative precision; and, when this has been done, the working procedure used for Pearson's method of components can then be adopted as before. Since lack of 'precision' is now assumed to be produced by the presence of 'specificity,' the calculations are circular, and have to proceed by successive approximation. We still postulate only as many factors as are statistically significant; and we are now enabled, with this minimum number of factors, to secure a far more plausible

¹ A similar problem arises in the physical sciences. Jeffreys (1931, p. 40) takes by way of illustration an experiment which yields seven correlated values for (i) the lapse of time, t , and (ii) the distance x travelled in that time by a falling body such as Newton's apple or one of Galileo's bullets. Were we to insist on a 'full analysis' in accordance with Dr. Warburton's principles, we could find a polynomial of the form $x = a_1 + a_2t + a_3t^2 + \dots + a_7t^6$, which would yield a 'perfect fit' for the observed values. But the principle of parsimony demands that, if possible, we should first postulate only one unknown constant, not seven, and seek the closest possible fit by applying the method of least squares. We might, of course, then discover that the discrepancies were so large that we were compelled to postulate a further factor—such as the resistance of the air; if not, we should be content to regard them as seven errors of observation. In this way, the physicist seeks to express the law of gravity in terms of the simplest possible formula, and the formula itself will be 'overdetermined by the data.' As Jeffreys puts it, "simplicity is a better guarantee of probability than accuracy of fit."

² Since with this procedure no allowance is taken of error, the common use of the word 'estimating' is hardly consistent with the technical definition usually given for the term in statistics.

fit¹ both to the observed correlations and to the observed measurements than either Pearson's procedure or Warburton's procedure would yield with the same number of factors. Moreover, it is now much easier to derive rigorous tests of significance. The practical drawback lies in the additional labour thus entailed. But, as electronic computers come into wider use, that criticism will lose much of its force, since the procedure is far better suited to such mechanical devices than the kind of rotation that Warburton or Thurstone would have us employ.

No doubt, a good many writers on mathematical statistics would nowadays be inclined to question a reliance on the principle of least squares. The most obvious way to surmount this objection, here as elsewhere, would be to invoke the principle of maximum likelihood. When applied to the initial data, the principle seems to break down.² But, when applied to the correlation matrix, it appears to lead to virtually the same formulæ as those reached by applying the 'correction for specificity.'

VI. THE SUBSTITUTION OF NEAR SPECIFICS

The last of the advantages that Dr. Warburton claims for his procedure is that "the specific factors of an ordinary analysis would be replaced by near-specific factors." He does not offer any formal proof to show that this would be so. In the tables that we ourselves have subjected to a 'full analysis,' we have found that the specific factors tend to be replaced *either* by near-specifics *or* by equally balanced bipolar factors mainly or exclusively affecting two traits or tests only³; but occasionally one of the factors turns out to be 'specific' in the ordinary sense, i.e., it has zero saturations for all but a single test. As a matter of fact, Dr. Warburton's own table includes an instance of this kind, since his last factor is a pure specific. But in any case it is not easy to see what particular merits the 'near-specific' would possess over and above the ordinary specific. Even if we argue that the errors of different correlation coefficients often tend to be correlated, this procedure is not likely to indicate how, if at all, such errors overlap.

Perhaps, however, it will be fairer to look a little more closely at the actual results secured with the analysis Dr. Warburton has proposed. How exactly does his table differ from the earlier table which he quotes from Professor Vernon's unpublished memorandum? In the first place, the 'Scholastic' factor obtained by Dr. Warburton, unlike that obtained by Professor Vernon, has no positive saturations for the Abstractions test. For this the reason is clear. Dr. Warburton decided to begin his triangular analysis by assuming that the Abstractions test might provisionally be regarded as a perfect test for the first or general factor. In Table I, therefore, it has a saturation of 1.00. On rotation this figure is only reduced by a very slight amount, namely, to .981. Consequently, with the Abstractions test there are no residuals left to form the basis of a positive saturation with any other factor: all the rest are negative (see Table III). It would, however, be difficult to justify so high a saturation for the first factor. The value proposed obviously makes the hypothetical correlations deduced from this factor far too large; and they in turn produce the negative saturations in the

¹ It does not give the *closest* fit: that is obtained by applying weighted summation to the reduced correlation matrix without explicit correction. Algebraically this procedure in turn is equivalent to regarding the values in the correlation matrix as variances and covariances, and the specific factor variances as equal. I may add that, whenever the considerations set forth above are not sufficient to yield a unique solution for the diagonal values, the best supplementary assumption would seem to be that which supposes (in the absence of other knowledge) that the unreliability or specificity is much the same for every test.

² Cf. M. S. Bartlett, this *Journal*, III, p. 81. See also Kendall, *ibid.*, p. 72, and Bartlett's comments, *ibid.*, p. 86. As applied by Lawley (1943) it seems to assume that what the physicist would call 'systematic errors' can be treated in the same way as, or at least pooled together with, 'accidental errors.'

³ Cf. this *Journal*, II, pp. 109-112, esp. Tables III, IV, and V. With weighted summation zeros are only likely to arise when certain artificial relations obtain between the correlations. If, with Dr. Warburton, we substitute simple summation for weighted, zeros will arise far more frequently (cf. *loc. cit.*, Table IV); the alternatives mentioned in the text then depend very largely on the way we choose to reflect the signs of the residuals.

later factors. Were these supplementary saturations genuinely non-significant, i.e., due to chance, we should expect them, not to be negative in every case, but to vary more or less equally on either side of zero.

The second difference arises from the diminished number of specific factors. Professor Vernon assumes eight; Dr. Warburton obtains only four. It is true that he describes his factors as 'near-specifics.' But, as we have seen, the last is perfectly specific; and the supplementary saturations of the other three are all non-significant. Indeed, with one exception the first figure is zero in every case.

Have we then any real grounds for assuming that, whereas Dictation, Electrical Information, Squares, and the Matrix test involve specific factors, the other four tests do not? The analyses carried out by Banks agree in finding large specific factors in the same four tests; but they also indicate, like the investigations of other factorists, the presence of an even larger specific factor for Arithmetic. And on more general grounds we should surely expect specific factors for Memory, Abstractions, and Mechanical Information (particularly if there is one for Electrical Information).

There can, however, be little doubt that the peculiar incidence of the four specifics (or 'near-specifics') is due solely to the way in which the rotation was effected. We have tried several alternative methods of rotating Dr. Warburton's triangular matrix; and the results show that as good a fit can be obtained by assigning the 'near-specifics' to an altogether different selection of tests. With the ordinary group factor method (where overlapping factors are derived by arithmetic rotation from a bipolar matrix of the usual type) the mode of rotation is fixed objectively by the pattern of signs; and the results are unique. With Dr. Warburton's method, on the other hand, both the triangular analysis and the rotation allow ample scope for subjective choice; and he offers no cogent reason for preferring the selection illustrated in Table III to any of the other possibilities.

Finally, the method Dr. Warburton recommends is decidedly laborious. We have first to calculate eight summational factors; next to rotate these factors, two or three at a time; then to compute the resulting 8×8 orthogonal rotation matrix; and, last of all, to deduce the rotated factor loadings by matrix multiplication. Nevertheless, in the investigation which he offers as typical, the outcome remains the same in nearly all essentials as that which is reached by the short and simple method of direct group factor analysis. The two most conspicuous changes, so far from constituting a manifest improvement, are in our view exceedingly questionable: judged against the background of previous enquiries, they appear, if anything, to distort rather than to improve the true picture.

However, it would scarcely be fair to condemn the proposed procedure after examining its application to just one set of correlations. The interesting suggestions which Dr. Warburton has put forward undoubtedly deserve further exploration; and in any case we may be grateful to him for the instructive questions he has raised in the course of his arguments.

VII. SUMMARY AND CONCLUSIONS

1. Expressed in terms of the number of factors, the principle of parsimony, taken by itself, requires us to find the minimum number of significant common factors needed to fit an observed set of measurements.

2. When the measurements relate to psychological traits, we are compelled by our antecedent knowledge of such data to allow for 'unique' or 'specific' factors peculiar to each test. This further condition may be met by seeking a satisfactory fit, not to the initial measurements, but to the table of observed covariances (or observed correlations if the measurements are in standard measure). We can therefore modify our original formulation, and say that the principle of parsimony requires us to find the minimum number of significant common factors which will fit the covariances observed within the usual margin of error. This implies that the leading diagonal of the covariance matrix will contain only that proportion of the several variances that is attributable to these common factors, i.e., it will contain reduced self-correlations, if the measurements have been standardized.

3. With this interpretation, Thurstone's statement that the reduced self-correlations (or 'communalities') must be so determined that the number of factors is equal to the rank of the side correlations appears erroneous. What we have to determine is not the smallest number of factors that will maintain the rank, but the smallest number of *significant* common factors required to render the residuals non-significant.

4. It would, however, be better still to apply the principle of parsimony to the number, not of significant common *factors*, but of significant factor *saturations*. If we assume that a set of summational (general-plus-bipolar) factors is required, the result, as a rule, will

be the same as before. But, if we are not restricted to this particular type of factor matrix, the principle will often indicate that a matrix of group factors or of subdivided factors is preferable, for then the number of saturations may prove to be smaller even when the number of factors is not.

5. Since, in any given case, the introduction of specific factors is based, not on a tentative assumption, but on antecedent knowledge, such factors cannot themselves be regarded as hypothetical entities whose presence has to be confirmed or refuted in the course of the enquiry. Hence, the principle of parsimony does not require us to minimize the number of specific factors—much less to reject them altogether (as we should in a full factor analysis carried out in accordance with Pearson's original procedure).

6. The method of full analysis, with subsequent rotation, recommended by Dr. Warburton and other writers, is inconsistent with the known characteristics of psychological data, where both specific factors and unreliability are nearly always large. Moreover, the working procedure would be needlessly laborious, and the results obtained by no means unique.

7. If the purpose of the calculations is purely arithmetical, namely, to transform the numerical data into some precisely equivalent but more convenient form, then the original method of principal components may be justified. If, however, as is more usual in psychological factorization, the purpose is statistical, namely, to determine the minimum number of hypothetical factors warranted by the data and their sampling errors, then a full analysis will no longer provide an appropriate answer. The technical objections to the ordinary procedures for determining the size of the specific factors can be overcome by 'correcting for specificity' in accordance with the device adopted in the more general form of the method of least squares.

REFERENCES

1. Burt, C. (1909). 'Experimental Tests of General Intelligence.' *Brit. J. Psychol.*, III, 94-177.
2. Burt, C. (1915). 'General and Specific Factors underlying the Primary Emotions.' *Brit. Assoc. Ann. Rep.*, LXXXV, 694-6.
3. Burt, C. (1917). *Distribution and Relations of Educational Abilities*. London : P. S. King.
4. Jeffreys, H. (1931). *Scientific Inference*. Cambridge : Univ. Press.
5. Frisch, R. (1934). *Confluence Analysis*. Oslo : Universitetes Økonomiske Institut.
6. Hartog, P. J., Rhodes, E. C., and Burt, C. (1936). *The Marks of Examiners*. London : Macmillan.
7. Ledermann, W. (1937). 'On the Rank of the Reduced Correlation Matrix.' *Psychometrika*, II, 85-93.
8. Holzinger, K. J., and Harman, H. H. (1941). *Factor Analysis : A Synthesis of Factorial Methods*. Chicago : University Press.
9. Lawley, D. N. (1943). 'The Application of the Maximum Likelihood Method to Factor Analysis.' *Brit. J. Psychol.*, XXXIII, 172-5.
10. Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago : University Press.
11. Thomson, G. (1948). *The Factorial Analysis of Human Ability*. London : Univ. London Press.
12. Banks, C. (1949). 'Factor Analysis of Assessments for Army Recruits.' *Brit. J. Psychol., Stat. Sect.*, II, 76-89.
13. Burt, C. (1949a). 'Pearson's Method Modified to allow for Specific Factors.' *Brit. J. Psychol., Stat., Sect.*, II, 65-6.
14. Burt, C. (1949b). 'Alternative Methods of Factor Analysis.' *Ibid.*, II, 98-121.
15. Thomson, G. (1949). 'The Nature of the Mind's Factors.' *Brit. Assoc. Ann. Rep.*, VI, 267-76.
16. Kendall, M. G., *et. al.* (1950). 'Factor Analysis as a Statistical Technique.' *J. Roy. Statist. Soc., Ser. B*, XII, 60-73, 85-94.
17. Banks, C. (1954). 'The Factorial Analysis of Crop Productivity.' *J. Roy. Statist. Soc., Ser. B*, XVI (in the press).
18. Warburton, F. W. (1954). 'The Full Factor Analysis.' *Brit. J. Stat. Psychol.*, VII, 101-5.

NOTES AND CORRESPONDENCE

EARLY RESEARCHES ON GENERAL AND SPECIAL ABILITIES

To the Editor, British Journal of Statistical Psychology

SIR,—In several recent articles by Mr. Durant and others, dealing with the factorial structure of human abilities, reference is made to the conclusions drawn from two early investigations, namely, the elaborate research carried out by Simpson under Thorndike's guidance, which appears to have been the first factorial enquiry undertaken in the United States, and what is described as "the first large factorial enquiry undertaken in Britain with group tests"—a research by R. C. Moore under the guidance of Professor Burt. In both cases the final factorizations, it would seem, were undertaken by or for a committee on 'Mental Factors in Education' set up by the British Association for the Advancement of Science. But unfortunately the printed reports, at least those accessible in the United States, do not give the detailed tables. Are these still available?

The former enquiry would seem to be of special importance in the history of factor analysis, since Spearman's demonstration of his 'Proof that *G* and *S* Exist' is based primarily on Simpson's figures (*Abilities of Man*, 1925, Chap. X, pp. 146f.). He explains that Simpson's data were selected as being those of "an investigator who, more than any other, has shown himself antagonistic to the doctrine of the general factor, *g*" (*Brit. Assoc. Ann. Rep.*, 1925, Pres. Add., p. 3). He himself claims that Simpson's correlations "display a most striking agreement" with the two factor theory. Nevertheless, more recent writers in the *Brit. J. Stat. Psychol.* present yet a third interpretation of Simpson's results. Perhaps therefore Professor Burt or Mr. Durant could give us the evidence for their own conclusions in fuller detail.—Yours, etc., DAVID M. FRIEDMANN.

Reply.—B. R. Simpson's monograph on 'Correlations of Mental Abilities' was originally published as Vol. 53 of the *Contributions to Education* (issued by Teachers' College, Columbia University, in 1912). The research, he says, was "suggested and outlined by Professor E. L. Thorndike."

He starts by pointing out that "students of psychology, beginning with Galton, have been devising tests of mental capacities, both *special* and *general*." His own interest, he explains, lies primarily in the practical attempt to develop a scheme of tests which could be used to determine "the nature and amount of a person's mental capacity." But there is a theoretical problem to be settled first: are we justified in distinguishing 'general intelligence' from more specialized abilities? "Can we hope to find a means of classifying pupils in this way?"

He gives a brief review of previous work on mental tests from Binet and Wissler onwards. Burt's factorial study is described in some detail, since it furnished the nearest parallel to the scheme of levels which Thorndike proposed for verification: the same investigation also formed the starting point for Moore's research to which Dr. Friedmann's letter refers. Working in McDougall's laboratory at Oxford, Burt and Flugel had applied, on two successive occasions, a set of individual tests to two groups of children—(a) bright (pupils from a preparatory school) and (b) ordinary or dull (pupils from an elementary school). The tests were designed to cover both afferent (sensory) and efferent (motor) aspects of cognitive activity, and four main cognitive levels: I. Elementary Sensory Discrimination, II. Elementary Motor Processes, III. Complex Sensori-Motor Processes, IV. Association, V. Attention or Apperception.¹ The results obtained indicated (i) a large factor for general cognitive ability and (ii) several smaller group factors, corresponding to the main levels tested,

¹ C. Burt, 'Experimental Tests of Higher Mental Processes and their Relation to Intelligence,' *J. Exp. Pedag.*, I, 1911, pp. 93–112. The results of this investigation were the subject of some correspondence with Thorndike, who had already carried out several correlational researches. Thorndike explained that he preferred to recognize only three main levels: (i) 'sensitivity,' (ii) 'association,' (iii) 'dissociation'—the last corresponding to Burt's highest level (called 'attention or apperception' in the article, but subsequently regarded as essentially 'relational': for this level tests of 'opposites,' 'analogies,' and 'reasoning' were included in the original programme, but were actually applied only in the supplementary study with group tests, carried out in conjunction with Moore). Thorndike also doubted the attempt to use or modify Pearson's method of seeking 'lines of closest fit,' chiefly on the ground that mental measurements were not amenable to the same methods of treatment as physical—a criticism subsequently made by Spearman. Unless the factors were 'obvious to inspection,' he thought mathematical methods of analysis would be 'unconvincing.'

notably association, sensori-motor capacity, and sensory discrimination, i.e., *both general and special* abilities. The underlying scheme and the correlation tables (with and without corrections for unreliability) are reproduced by Simpson in full.

His own investigation follows a similar plan. He applies, individually and twice over, 14 mental tests representing six main levels or aspects to two groups of adults, namely, (a) 'bright,' consisting of 17 professors and advanced University students, and (b) 'average or dull,' consisting of 20 persons found in Salvation Army Homes or elsewhere. The observed correlations are corrected for unreliability and selection; and his inferences are based chiefly on a table of 'estimated true coefficients for people in general' (i.e., for 'persons representing a *random* sample of the whole population'). He agrees, however, that the 'gain in accuracy' resulting from these rather laborious corrections is open to doubt; and nearly all later writers who have discussed or re-analysed his data have preferred to work with his initial table of observed correlations obtained from the two groups pooled together. This is reproduced below¹ in Table I.

TABLE I. CORRELATIONS BETWEEN TESTS (Simpson)

Tests	1 C.S.	2 H.O.	3 R.F.	4 L.F.W.	5 M.P.	6 M.W.	7 G.F.	8 Er.	9 C.W.	10 Sc.	11 E.O.	12 Add.	13 E.L.	14 D.L.
1. C.S. ..	—	.98	.88	.78	.91	.94	.54	.62	.42	.55	.79	.71	.33	.25
2. H.O. ..	.98	—	.74	.73	.81	.84	.70	.64	.43	.52	.80	.79	.26	.25
3. R.F. ..	.88	.74	—	.69	.66	.71	.49	.51	.37	.44	.56	.47	.34	.28
4. L.F.W. ..	.78	.73	.69	—	.59	.73	.35	.39	.29	.36	.42	.39	.26	.09
5. M.P. ..	.91	.81	.66	.59	—	.82	.57	.55	.31	.54	.52	.53	.28	.19
6. M.W. ..	.94	.84	.71	.73	.82	—	.56	.55	.40	.53	.62	.49	.28	.21
7. G.F. ..	.54	.70	.49	.35	.57	.56	—	.73	.56	.54	.53	.45	.25	.25
8. Er. ..	.62	.64	.51	.39	.55	.55	.73	—	.59	.39	.57	.54	.25	.22
9. C.W. ..	.42	.43	.37	.29	.31	.40	.56	.59	—	.31	.29	.57	.21	.07
10. Sc. ..	.55	.52	.44	.36	.54	.53	.54	.39	.31	—	.45	.51	.19	.27
11. E.O. ..	.79	.80	.56	.42	.52	.62	.53	.57	.29	.45	—	.68	.38	.48
12. Add. ..	.71	.79	.47	.39	.53	.49	.45	.54	.57	.51	.68	—	.17	.25
13. E.L. ..	.33	.26	.34	.26	.28	.28	.25	.25	.21	.19	.38	.17	—	.24
14. D.L. ..	.25	.25	.28	.09	.19	.21	.25	.22	.07	.27	.48	.25	.24	—

Simpson arranges his tests in order of the total observed correlation for each, in accordance with the method suggested by Burt. But he does not use these sums to derive the factor-saturations or the expected ('hypothetical') correlations. Nor does he formally test the residual correlations for statistical significance. Indeed, he considers Burt's "resort to further manipulation of the data" as a little too speculative to be trusted. Instead he prefers to pick out by inspection those pairs or clusters of tests (e.g., the memory tests and the perceptuo-motor tests) which have much higher intercorrelations than their position in the 'hierarchy' would lead us to expect. He notes that, as a result of the 'special abilities' presumably underlying such clusters, the correlation table shows nothing like a regularly descending hierarchical order. "Thus it is quite evident that Spearman's theory is not in harmony with the facts we have secured."

His main conclusions are summed up as follows: "We find justification for the assumption that there is a something that may be called '*general* mental ability' or '*general* intelligence' and that on the other hand certain capacities are relatively *specialized*." The size of the total correlations indicates that, contrary to Spearman's declaration, '*sensory discrimination*' yields the poorest measure of general intelligence, and tests of '*abstract thinking*' the best. Simpson distinguishes six

¹ The table has been rearranged to show the clustering of high correlations due to several group factors. For the most part the names of the tests explain themselves. 'Completing Sentences' consists of an Ebbinghaus *Kombination* test; 'Completing Words' denotes a test in which the examinee is to add letters to syllables like 'ba, ca, da,' etc., to make a complete word. In 'Recognizing Forms' the examinee has first to study 25 small shapes and then mark on a larger sheet containing a wide variety of shapes those that he has just seen. In 'Geometrical Forms' he has to cross out certain shapes; in 'Erasures,' certain letters. In 'Learning Names' he is given one minute to study a column of arbitrary shapes against each of which a word is printed, to serve as the name or translation of the hieroglyph: he is then shown only the shapes, and required to write down the word associated with each. 'Scroll' is the familiar tracing test intended to measure 'motor control.'

special abilities, and criticizes Burt for not separating those activities which involve memory (in the sense of learning) from those which depend on associations already acquired. For the rest, as he observes, his conclusions agree pretty closely with the results of the Oxford experiment.

Simpson's work attracted considerable attention at the time. The doctrine of mental 'levels'—first put forward in this country, then accepted by Thorndike, and here confirmed by the experiments of one of Thorndike's own research students—appeared to be in marked contradiction with Spearman's 'two-factor theory,' since this flatly denied the existence of any common factors other than the general factor, *g*. Spearman immediately attacked it, first at a 'Symposium on Mental Tests' arranged by the Education Section of the British Association and then in a joint paper with Hart: in the latter he applied the intercolumnar criterion to Burt's tables in order to show that they really confirmed his 'unifocal theory,' not 'Thorndike's multifocal theory of faculties or levels' (*Brit. J. Psychol.*, V, 1912, esp. pp. 52-61). Accordingly, as secretaries of the British Association Committee on Factors (mentioned by Dr. Friedmann), Professor J. A. Green and Burt were asked to prepare a report reviewing investigations hitherto carried out on the subject, including those of Simpson, Brown, Moore, and others.

In analysing Simpson's data, two methods were employed; and the factor-saturations obtained by their means are reproduced in Table II.

TABLE II. FACTOR-SATURATIONS

Tests	(a) General and Bipolar Factors			(b) Basic and Group Factors						
	I	II	III	I'	II'	III'	IV'	V'	VI'	VII'
1. Completing Sentences ..	.980	.313	.079	.791	.604	.306	—	—	—	—
2. Harder Opposites ..	.952	.117	.071	.812	.427	.178	—	—	.462	—
3. Recognizing Forms ..	.790	.208	.102	.735	.346	.253	—	—	—	—
4. Learning Forms and Words	.680	.442	.027	.542	.552	.241	.197	—	—	—
5. Memory for Passages ..	.811	.252	— .135	.742	.441	—	.212	—	—	—
6. Memory for Words ..	.864	.315	— .144	.735	.430	—	.423	—	—	—
7. Geometrical Forms ..	.739	— .308	— .381	.737	—	—	—	.417	—	—
8. Erasures ..	.728	— .266	— .226	.696	—	—	—	.468	—	—
9. Completing Words ..	.527	— .283	— .188	.501	—	—	—	.689	—	—
10. Scroll Motor Test ..	.604	— .084	— .072	.652	—	—	—	—	—	—
11. Easy Opposites ..	.809	— .226	.462	.671	.248	—	—	—	.522	—
12. Addition ..	.716	— .165	.032	.693	.153	—	—	—	.397	—
13. Estimating Lengths ..	.363	— .060	.105	.374	—	—	—	—	—	.359
14. Discriminating Lengths..	.333	— .255	.268	.295	—	—	—	—	—	.359
Contribution to Variance (Per cent.)	53.6	6.4	4.2	40.4	10.3	1.8	1.9	6.2	4.6	1.8

(a) *Summational Analysis.* The first analysis was based on the 'simple summation' formula already used by Burt, viz. (in the usual notation) $r_{ag} = \frac{r_{ai}}{\sqrt{\sum \sum r_{ai}^2}}$. Three factors were extracted contributing 54 per cent., 6 per cent., and 4 per cent. respectively to the total variance.

I. It will be seen that, in Simpson's investigation, the first or general factor appears to contribute rather more, and the supplementary factors rather less, than the proportions usually reported in more recent enquiries. These peculiarities, and the exceptionally high correlations that he had obtained, undoubtedly arise from the fact that, as a result of his experimental plan, the total group of persons tested contained a disproportionate number of unusually intelligent and unusually unintelligent persons. This, however, only enlarges the saturations and does not affect their relative size. The figures we obtained for this factor were fully in keeping with Simpson's main conclusion, namely, that the tests of 'abstract thinking' (completing sentences, hard opposites, and easy opposites)

provided the best tests of general ability (saturation over 0.8) and the two tests of discrimination the poorest (both saturations under 0.4).

II. The second factor is bipolar, and divides the whole series of tests into two main groups—(A) six belonging to the higher mental levels and (B) eight belonging to the lower. Roughly speaking, the former are tests of a more abstract or intellectual type, the latter tests of a more concrete, practical, or mechanical type.

III. The third factor subdivides each of the preceding groups into two subsections, making four levels in all. (A1) The first and highest level consists mainly of apperceptive or relational tests (tests of 'selective thinking,' as Simpson prefers to call them); (A2) the second consists of the tests of memory (in the sense of short-distance learning); (B1) the third, of the perceptual tests; while (B2) the fourth and lowest comprises the most elementary and mechanical tests.

With the chi-squared criterion the second factor is just significant ($P = .04$) and the third hardly significant ($P = .06$). The 'Scroll' test, it will be observed, has saturations of approximately zero for both the supplementary factors; and so stands apart.

(b) *Group Factor Analysis.* To obtain factors which would correspond more directly with the 'special abilities' referred to by Simpson and denied by Spearman, an alternative method of analysis was used which would yield group factors with positive saturations only instead of dichotomous factors with negative as well as positive saturations. The procedure adopted was a cruder form of what is now known as group factor analysis: it involved (a) first calculating a general factor from those observed correlations showing no signs of supplementary factors, and then (b) calculating the saturations for the group factors from the clusters of residuals. The same summation formula was used, but applied to what would now be called submatrices instead of to the table as a whole. The results are shown in Table II(b).

Here the first factor is a 'basic factor,' which might be identified with 'general intelligence' as ordinarily understood. 'Completing Sentences' and 'Hard Opposites' still yield the highest saturations; but 'Easy Opposites' now appears less satisfactory than before. For the rest the chief conclusions remain the same. There is a large group factor covering all the tests involving 'associative' processes, in the broadest sense of the word: even 'Easy Opposites' and 'Addition' show small saturations for this factor. It would seem to be the same as the 'intellectual factor' which so conspicuously figures in later factorizations of cognitive tests. Vernon prefers to describe it as 'v-ed' (verbal-educational). But with children it seems to represent rather a capacity for successful school education than the actual effects of school education; nor is it by any means purely verbal: arithmetical tests (when included) usually seem to come under the same head. This large group factor corresponds with the positive half (A) of the first bipolar in the previous analysis: but no group factor is discernible to correspond with the negative half (B), or, if there is, its saturations are too low to be significant.

There is a smaller group (corresponding to section A1 in the third factor of the preceding analysis) which is presumably a factor for the highest apperceptive and relational processes, and another (corresponding to section A2) which enters into the tests of memory and of learning words or shapes (short distance learning). The remaining factors appear to group together first the three perceptual tests, then the two tests involving what Simpson calls 'mechanical association' (long-distance learning), and finally the two tests of sensory discrimination. These correspond to sections B1 and B2—the latter being here subdivided.

Except for the second, all the foregoing factors are noted by Simpson, though the evidence that he himself submits—often, it would seem, resting as much on preconceptions about the nature of his tests as on his actual figures—is not altogether conclusive. On the other hand, he postulates a small specific factor (or 'capacity') for 'Motor Control,' which, he believes, is tested by the Scroll test. Since this test (as the bipolar analysis shows) is the sole representative of its group, no group factor of this kind could be demonstrated by the methods here used: indeed, with this particular battery the factor is really identical with a 'specific' factor in the narrow sense, i.e., a 'unique' factor peculiar to this particular test.

When it appeared, Simpson's monograph—a volume of 122 pages—was the first book devoted wholly to a factorial problem. Moreover, it was taken to represent the view to which Thorndike had been led as a result of his examination of Spearman's theory. That, I imagine, is why Spearman himself laid so much stress on it in his own contributions. In addition to the two references cited in Dr. Friedmann's letter, he has discussed it at some length in an article in the *Psychological Review* (XXI, pp. 101f.) and more briefly in his recent book on *Human Ability* (with Wynn Jones, 1950), p. 138.

Dr. Friedmann enquires how it is that such opposite conclusions have been drawn from Simpson's data. There were two reasons. First, Simpson trusted to a simple inspection of his correlation table to discover the clusters of augmented correlations due to group factors, and rejected the use of more objective devices. Secondly, the number of persons tested was extremely small (that indeed was the weak point of most researches in the days of individual testing); hence without some further 'mathematical manipulation of the data' it was impossible to be sure that the increased size of

certain correlations might not after all be due to sampling. Both these defects were noted by Spearman; and very similar criticisms were subsequently made by F. N. Freeman.¹

In applying tests of statistical significance to the residual correlations, Burt, Green, and Moore, like most other investigators, were content to accept as significant deviations that were three times the probable error. Spearman, on the other hand, assumed that no deviations were worth considering unless they were *five* times the probable error. Even so, the way he words his inferences suggests that he has proved and disproved far more than is actually the case. He tells us, for example, that he singled out Simpson's research because it provided "the strongest case" against the two factor theory, so that, if Simpson's arguments were refuted, there would be little ground left for opposition. But that is scarcely correct. Simpson himself observes that, *prima facie*, Burt's figures "even *more strongly* contradict Spearman's original theory of a hierarchical arrangement than do our results spoken of above." And between 1909 and 1925 many other correlation tables had been published by workers in this country which were based on much larger samples of persons and often on a larger number of tests.² It is therefore unfortunate that Spearman did not apply his tetrad-difference criterion to these other tables as well.³

But secondly, Spearman's incidental comments (quoted by Dr. Friedmann) suggest that Simpson himself had discerned no signs of a general factor. In point of fact, Simpson fully recognizes its presence; and he makes it quite clear that his main divergence from Spearman arises, not over the general factor (which both accept), but over the special or supplementary factors (which Simpson accepts and Spearman denies): it is these that "break Spearman's alleged hierarchical order." Spearman rejoins that the breach of the hierarchy could be explained perfectly well by sampling errors. But even so that would merely mean that the numbers tested were too few to afford a conclusive demonstration of any supplementary factors. Strange to say, in a footnote Spearman himself admits that there *were* some tetrad differences which exceeded even $5 \times \text{p.e.}$ and "therefore cannot well be attributed merely to sampling errors." Consequently he is hardly justified in claiming in the text that the agreement between his own theory and his analysis of Simpson's data is "most striking."

Spearman goes on to contrast the results of analysing Simpson's correlation table for *mental* measurements and the well-known Pearson-Macdonell correlation table for *physical* measurements, and prints graphs of the tetrad differences side by side. With the table for physical measurements the criterion shows unmistakable evidence of the presence of supplementary factors. But the reason for the wide difference in the range of the probable errors for the two cases is simply that the Pearson-Macdonell table was based on over 2,000 cases and the Simpson table on only 37.

In our view, however, the fundamental issue is wrongly stated. The question to ask is not: are there any supplementary factors besides the single general factor and the inevitable specifics? Take sufficiently large numbers, and you are almost certain to find that a single general factor will not suffice for an adequate fit. The real question is: how much do these supplementary factors contribute to the total variance as compared with the general factor? Burt's analysis of the Pearson-Macdonell data showed a general factor accounting for about 50 per cent. of the variance and a succession of

¹ "Simpson," he says, "presents his results in the form of a table similar to that of Burt. The order of the tests is based on the average intercorrelation, and bears out the same conclusion drawn from Burt's. . . . But, if Spearman's theory is correct, capacities do *not* fall into groups. . . . Doubts about the validity of Simpson's classification are deepened by the fact that in many cases the correlations between traits which fall in *different* classes are closer than many of the correlations between traits in the *same* class. This forces on us the question whether a classification of special abilities must not follow some other line" (*Mental Tests*, 1926, pp. 77-9: Freeman does not indicate what 'other line' he has in mind, but his later references suggest that he is thinking of some mathematical form of group factor analysis).

² The research carried out by Mr. R. C. Moore, to which Dr. Friedmann refers, included 22 tests applied to 130 children in Wallasey schools (*J. Exp. Ped.*, 1, 1912, p. 375). Mr. Moore, it is hoped, will himself publish a fuller account of his analysis in a forthcoming issue of this *Journal*. Hitherto all researches had been carried out with individual tests and were therefore confined to small groups. Burt had tentatively used group tests in one of his own earlier studies; but Spearman had argued that data from small groups tested individually would, when 'corrected' for unreliability, prove far more trustworthy than data from larger groups "tested in the mass." The main purpose of Mr. Moore's enquiry was to demonstrate that group tests would give results that were equally reliable and equally valid.

³ In *Abilities of Man* he also includes two analyses for small batteries consisting of intelligence-tests only; but naturally this is not the kind of data in which we should expect to find evidence of special abilities. No doubt, Spearman's difficulty was the vast amount of labour entailed in applying his criterion: with Simpson's table 3,003 tetrad differences had to be calculated by the laboratory technician, Mr. George Raper.

bipolar factors accounting for 12, 7, and 5 per cent. ; these are very much the same proportions as were found in his 1909-1914 investigations on mental tests. Simpson's data certainly show a rather larger contribution for the general factor (nearly 54 per cent.) and rather less for the group factors (the next largest factor accounts for only 6 per cent.). But this slight difference arises from the fact (which Spearman does not explain) that the persons tested by Simpson were not intended to represent a random sample of the population, but formed a composite group "representing as far as possible two extremes of general intelligence." That would naturally increase the prominence of the general factor at the expense of the more specialized.

We think, therefore, that it is quite impossible to accept Spearman's claim that Simpson's results really confirm the 'two factor theory of g and s ,' and display "one of the most striking agreements ever recorded in psychology." On the contrary, they show a far closer agreement with the theory Spearman believed he was disproving—the theory that in addition to ' g and s ' there was a third series of factors, namely, group factors of both broad and narrow range.

Unfortunately in America interest in factorial methods lapsed almost entirely until it was revived afresh, sixteen years later, by Kelley's book, *Crossroads in the Mind of Man* (1928). Kelley favoured a scheme of "general, group, and specific" factors, while insisting (as British writers had done) that the distinctions were merely relative : he recognizes, in addition to a 'general factor,' a number of special factors both for 'content' (verbal, numerical, etc.) and for 'process' (sensory, motor, memory, etc.), very similar to the lists published by British workers (cf. Burt, *Measurement of Mental Capacities*, 1926). All this was fully in keeping with the views developed by Thorndike and Simpson, though, curiously enough, his 'bibliography' makes no mention of Simpson's research. The subsequent attempt to produce schemes of 'primary abilities' consisting of group factors only (with no general factor) appears to mark a new starting-point in American factor theory ; and, as a result, the earlier work of Thorndike and his pupils seems to have been almost forgotten, even by those who have written brief histories of the subject.

C. BURT
P. DURANT.

JOURNAL COMMITTEE

The Publications Committee of the British Psychological Society has recommended the appointment of a journal committee for each of the Society's publications. Accordingly at its meeting on May 1st, 1954, the Council appointed the following as members of the Journal Committee for the *British Journal of Statistical Psychology* (in addition to the Editor and Assistant Editors) :

Mr. Brian Foss,
Dr. Max Hamilton,
Mr. Alistair Heron,

Dr. F. W. Warburton,
Dr. John Sutherland,
Dr. R. H. Thouless.

The first meeting was held on July 16th. The chief question discussed was the financial position of the *Journal* and the Council's proposal that it should be printed by the Aberdeen University Press. The Committee recommended that this and other suggestions should be more fully examined before action was taken.

A second meeting was held on September 9th ; and, as requested by the President, the Committee drafted a detailed report, on the basis of information gathered in the interim, for submission to Council. The Editor expressed the hope that it would be possible to introduce more generous spacing and dispense with small type in the text. Owing to improved facilities recently provided by the Monotype Corporation for mathematical work, he thought it might now be cheaper to change from Times Roman (the type face hitherto adopted) to Modern Series No. 7 (the style preferred by *Biometrika*, *Psychometrika*, and most mathematical journals). From estimates received from various London firms accustomed to printing mathematical work (including the present printers) it would seem that, with these changes, cost of production could be very greatly reduced.¹ The Editor also expressed a wish to receive contributions dealing with a wider range of subjects, including notes, questions, or abstracts of theses from younger members of the Society. The Committee would be grateful to receive criticisms and suggestions from readers, particularly in regard to the foregoing proposals.

Charlotte Banks. Secretary.

¹ At its meeting on October 2, Council decided to reject the recommendations of the Committee, and reaffirmed its former resolution, namely, that the *Journal* should in future be published by the Society and printed by the Aberdeen University Press. It was agreed that questions of type-face should be referred to a Sub-committee of the new Publications Committee when appointed.

The Editor would like to express his deep appreciation of the constant help given during the past eight years by the University of London Press, and to thank them for their advice and aid in reducing the heavy deficit.

BOOK REVIEWS

The Geometry of Mental Measurement. By G. H. Thomson. London : University of London Press. 1954. Pp. 60. 6s. 6d.

This little book is based on three lectures delivered at the University of London in 1953. Bivariate correlation forms the starting point, and leads to a discussion of partial and multiple correlation and regression ; the book is principally concerned, however, with the concepts of factor analysis. The clarity of the exposition is greatly aided by simple pictorial examples in which the reader is invited to visualize dimensions in terms of rooms and table-tops. Most of the diagrams are excellent ; but in one or two of the more complex the rather large lettering at times makes identification difficult.

The book will be found instructive by advanced students as well as by beginners. In the closing sections the consequences of using communalities are examined from a geometrical standpoint, and the result is highly illuminating. Unfortunately the discussion is so phrased as to suggest that uncorrelated factor estimates can invariably be realized as actual numerical quantities. It is true that this is qualified by the statement that "we cannot allot these estimates to individual men, so they are not of practical use"; but the implications might be found misleading. In the same section the reader learns that another writer (Kestelman, this *Journal*, V, pp. 1-6) has shown "how the values of the uncorrelated factor-estimates can be *calculated*" (my italics) when communalities are used. Kestelman, however, merely proved that such estimates *must exist* ; he did not give a process for their calculation. In spite of these minor defects the book can be heartily recommended to those whose acquaintance with factor analysis is restricted to its arithmetical techniques or to an algebraic approach. A page or two giving references for further reading and an index of subjects would probably be welcomed by the novice who makes this book his primer.

Arthur Summerfield

Psychological Testing. By Anne Anastasi. New York : The Macmillan Company, 1954. Pp. xvi + 680. \$6.75.

Professor Anastasi's main aim is "to provide an introduction to the principles of psychological testing and acquaint the reader with the major types of tests in current use." Her book is divided into four 'parts.' The first discusses with admirable clarity the origin, nature, and purpose of mental tests, and the problems of reliability, validity, and interpretation ; the three remaining sections describe in turn the commoner methods of testing general intelligence, specific aptitudes, and 'personality characteristics.' As the book is designed for practical workers, the author rightly decided that, instead of the usual 'statistical chapter,' it would be better to introduce statistical concepts as they are needed, explaining each in an appropriate context. British readers may at times be tempted to criticize the context chosen. Writers in this country seem usually to start off with the results of factor analysis, as indicating the chief distinguishable aspects of human personality which the psychologist is called upon to test. In the present volume factor analysis is not mentioned until we reach Part III (the differential testing of special abilities). Consequently, the evidence for accepting the concept of a 'general cognitive ability,' distinct from 'special or group factors,' is omitted. Dr. Anastasi, however, explains that "intelligence tests are now more properly described as general classification tests." Even so, however, if a *general* classification is to claim validity, that surely can only be done by demonstrating the presence of a *general* factor. No doubt factorial analyses by themselves can do no more than confirm such a hypothesis ; but its advocates usually contend that it is supported by experimental, neurological, and biological evidence as well.

Some critics too will also question the historical account of early British work. We are told that the factorial method was "initiated by Charles Spearman," and that later "American investigators, such as Kelley and Thurstone, proposed a number of group factors, rather than a single *g* factor." Spearman, however, claimed, not to initiate, but merely to correct and simplify the earlier multifactor techniques developed by Pearson and his followers. And the distinction between 'general ability' and 'special abilities' was due not to Spearman but to Galton.

These, however, are minor points ; and every reader will readily acknowledge the clear and scholarly way in which the author has discussed both problems and procedures.

Florence Mitchell

The Structure of Human Personality. By H. J. Eysenck. London : Methuen & Co. 1953. Pp. xx - 348. £1 17s. 6d.

During the past few years a succession of publications has appeared which attempts to describe, on the basis of factorial results, the structural organization of individual personality. Burt started with his papers on *The Structure of the Mind*; Vernon followed with *The Structure of Human Abilities*; Cattell's book on *Personality* included chapters on its 'intrinsic structure'; and in his latest publication Dr. Eysenck now summarizes his own views of *The Structure of Human Personality*. In discussing the author's previous work, an earlier critic very naturally asked how far the conceptions set forth agreed or disagreed with those of previous writers. In his reply (this *Journal*, VI, p. 45), Dr. Eysenck explained that the preceding volume was not "an historical treatise" and that a "systematic attempt to cover the ground" would "appear as a separate publication under the title *The Structure of Human Personality*." Here, therefore, we have his promised answer.

The book begins with a short introduction on 'theories of personality organization.' It is, however, somewhat disappointing, since it virtually ends with Jung and Kretschmer. In the remainder of the work, as in the opening chapter, the comparisons are chiefly limited to German and American authors; and for the rest the problems, methods, and contentions are much the same as before. As Professor Drever has remarked, a reader of Dr. Eysenck's successive publications at times "gets the feeling that he is watching the parade of a stage army: familiar figures pass and repass, sometimes with beards and sometimes without."

In the present volume, however, there is one important change. A "distinct preference" is now expressed for Professor Thurstone's techniques, which in his earlier writings Dr. Eysenck had severely criticized and rejected. Nevertheless, in practice, he still seems to prefer theories and procedures that are distinctive of British work. He starts by formulating a theory of four factors—general, group, specific, and error—in words that are almost a verbatim quotation of Burt's theory; and then proceeds to develop a view of 'the hierarchical structure' of personality, involving 'subdivided factors,' which (together with the tree-diagram printed to illustrate it) also closely resembles that described by Burt. Now these are the two main features in which transatlantic writers differ from British: as Vernon puts it, "Thurstone and other factorists in America from 1938 on have opposed the notion of a 'general factor' and a 'hierarchy.'" Again, on a later page Dr. Eysenck complains that "Cattell's method of analysis, taken over from Thurstone, exaggerates the size of the communalities and consequently the number of the factors," and moreover "results in oblique factors"—a criticism to which the present reviewer would heartily assent.

The outcome of the work is summarized in four main conclusions. First, "the agreement between different experimenters is more marked than might have been expected from their very vocal disagreements." Secondly, "work in the field of personality should embrace personality in all its aspects." Yet here again Dr. Eysenck does not always practise what he preaches. The simultaneous study of cases by clinical methods has always been typical of British factorists as contrasted with American: yet the clinical approach is here entirely ignored. As another reviewer has already observed (*Brit. J. Med. Psychol.*, XXVII, p. 261), Dr. Eysenck tacitly assumes that one can factorize and interpret figures for a group of people in a mental hospital, "without ever seeing the subjects—that is, the persons or the patients." Thirdly, there is a need for more intensive studies "integrating factorial with experimental work on the individual tests"; on this every reader will agree. Finally, instead of trusting to a "blind empiricism," as other factorists are said to have done, "the technique of factor analysis should be used in the service of a hypothetico-deductive methodology."

The style of the book is readable and pungent: some may find it a little too pungent (as when Jaensch's work is said to suffer from a "confused semantic diarrhoea"). Occasionally there are lapses into a German idiom (e.g., "In the years that followed a number *have been* published," and "1,540 correlations were run between 56 girls"); and there are numerous signs of hasty writing (as when J. G. Beebe-Center appears as "Beebe-Centre," Godfrey Thomson as "Thompson," and Sahai's work is repeatedly confused with Sanai's). But even those who are at times disposed to disagree with the arguments or inferences set forth will find much that is stimulating and suggestive.

R. L. Cohen

The National Readership Survey : 1954 Edition. The Institute of Incorporated Practitioners in Advertising, 1954. Pp. 115. £10 10s.

The object of the investigation reported in this volume was to secure a survey of reading habits among the different sections of the population of Great Britain. An endeavour has been made to base the whole enquiry on the most up-to-date techniques, and to cover, so far as was practicable, both the national and the provincial press. The 15,000 persons interviewed were intended to form "a stratified random sample of individuals," classified according to age, sex, region, and socio-economic status; and the methods for securing such a sample were devised in consultation with Professor

Maurice Kendall. The questionnaire involved both "a visual and a verbal aided recall technique," and was compiled with the assistance of Mr. Alec Rodger, Reader in Psychology in the University of London. The bulk of the book consists of 45 tables printed on oblong foolscap pages ($13\frac{1}{2} \times 17$ in.). The typography is somewhat unusual; and the publishers follow the excellent and growing practice of stating on the verso of the title-page how the book is set up; the text is set in Monotype Bodoni; the tables are in Monotype Gill Sans, and look as excellent as if they had been composed by hand.

The enquiry was undertaken primarily with a view to meeting the needs of advertisers and their advertising agents. Burt, however, has shown how data of this kind may be of the greatest value to the psychologist; and, in a recent discussion of the adult distribution of ability,¹ has already used earlier percentages for newspaper reading and the like to throw light on the intellectual interests and attainments of various social groups. The more extensive and reliable figures published in this report, together with supplementary tables on cinema-going, television, drinking, smoking, etc., will provide a mine of curious facts for the social psychologist, and should now permit still more detailed and accurate conclusions to be drawn.

The planning of the survey, the field work itself, and the statistical analysis may well serve as models to future investigators; and we are tempted to suggest that a further classification of the original data might well be undertaken from a psychological standpoint.

Charlotte Banks

INDEX TO VOLUME VII

INDEX OF AUTHORS

<i>Authors</i>	<i>Titles</i>	<i>Page</i>
Banks, C., and Burt, Cyril	The Reduced Correlation Matrix	107-117
Barlow, J. A., and Burt, Cyril	The Identification of Factors from Different Experiments	52-56
Burt, Cyril	The Sign Pattern of Factor-Matrices	15-29
Burt, Cyril	A Reliability Coefficient based on Analysis of Variance	61
Durant, P., and Burt, Cyril	Early Researches on General and Special Abilities	119-124
Gabriel, K. R.	The Simplex Structure of the Progressive Matrices Test	9-14
Harper, F. N.	Spearman's Form of Criterion Analysis	57-61
Highmore, G., and Taylor, W. R.	A Factorial Analysis of Athletic Ability	1-8
Howie, D.	A Comparison of Two Methods of Factorizing Test Data	31-36
Jonckheere, A. R.	A Test of Significance for the Relation between m Rankings and k Ranked Categories	93-100
Lawley, D. N., and Swanson, Z.	Tests of Significance in a Factor Analysis of Artificial Data	75-79
Millican, G. D., and Vernon, P. E.	A Further Study of the Reliability of English Essays	65-74
Neuhaus, J. O., and Wrigley, C.	The Quartimax Method	81-91
Stuart, A.	The Correlation between Variate-Values and Ranks in Samples from a Continuous Distribution	37-44
Stuart, A.	Limit Distributions for Total Rank Values	50-51
Swanson, Z., and Lawley, D. N.	Tests of Significance in a Factor Analysis of Artificial Data	75-79
Taylor, W. R., and Highmore, G.	A Factorial Analysis of Athletic Ability	1-8
Vernon, P. E., and Millican, G. D.	A Further Study of the Reliability of English Essays	65-74
Warburton, F. W.	The Full Factor Analysis	101-106
Whitfield, J. W.	The Distribution of the Difference in Total Rank Value for Two Particular Objects in m Rankings of n Objects	45-49
Wrigley, C., and Neuhaus, J. O.	The Quartimax Method	81-91

¹ *The Contribution of Psychology to Social Problems* (Oxford University Press), pp. 72f.

INDEX OF SUBJECTS

<i>Titles</i>	<i>Authors</i>	<i>Page</i>
Athletic Ability, a Factorial Analysis of	Highmore, G., and Taylor, W. R. . .	1-8
Criterion Analysis, Spearman's Form of	Harper, F. N. . .	57-61
Factor Analysis of Artificial Data, Tests of Significance in a	Lawley, D. N., and Swanson, Z. . .	75-79
Full Factor Analysis, the	Warburton, F. W. .	101-106
General and Special Abilities, Early Researches on	Durant, P., and Burt, Cyril . . .	119-124
Identification of Factors from Different Experiments, the	Barlow, J. A., and Burt, Cyril . . .	52-56
Quartimax Method, the	Neuhaus, J. O., and Wrigley, C. . .	81-91
Reduced Correlation Matrix, the	Banks, C., and Burt, Cyril . . .	107-117
Relation between m Rankings and k Ranked Categories, a Test of Significance for the	Jonckheere, A. R. .	93-100
Reliability Coefficient based on Analysis of Variance, A	Burt, Cyril . . .	61
Reliability of English Essays, a Further Study of the	Millican, G. D., and Vernon, P. E. . .	65-74
Sign Pattern of Factor-Matrices, The	Burt, Cyril . . .	15-29
Simplex Structure of the Progressive Matrices Test, the	Gabriel, K. R. . .	9-14
Total Rank Values, Limit Distributions for	Stuart, A. . . .	50-51
Total Rank Value for Two Particular Objects in m Rankings of n Objects, the Distribution of the Difference in	Whitfield, J. W. . .	45-49
Two Methods of Factorizing Test Data, a Comparison of	Howie, D. . . .	31-36
Variate-Values and Ranks in Samples from a Continuous Distribution, the Correlation between	Stuart, A. . . .	37-44

BOOKS REVIEWED

<i>Authors</i>	<i>Titles</i>	<i>Page</i>
Anastasi, A.	Psychological Testing	125
Brookes, B. C.	Notes on the Teaching of Statistics in Schools	64
Eysenck, H. J.	The Structure of Human Personality	126
Institute of Incorporated Practitioners in Advertising	The National Readership Survey : 1954 Edition	126-127
McKinsey, J. C. C.	Introduction to the Theory of Games	63-64
Notcutt, B.	The Psychology of Personality	63
Stephenson, W.	The Study of Behavior : Q-Technique and its Methodology	62-63
Thomson, G. H.	The Geometry of Mental Measurement	125

THE
BRITISH JOURNAL
OF
STATISTICAL
PSYCHOLOGY

EDITED BY
CYRIL BURT

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ARTHUR SUMMERFIELD

AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN	L. S. PENROSE
M. S. BARTLETT	J. FRASER ROBERTS
W. G. EMMETT	A. RODGER
M. G. KENDALL	W. STEPHENSON
D. N. LAWLEY	P. E. VERNON
E. S. PEARSON	J. W. WHITFIELD

Managing Editor : H. G. MAULE

UNIVERSITY OF LONDON PRESS LTD.

WARWICK SQUARE, LONDON, E.C.4.

Price Fifteen Shillings

CONTENTS

	Page
HIGHMORE, G., AND TAYLOR, W. R.	A Factorial Analysis of Athletic Ability 1
GABRIEL, K. R.	The Simplex Structure of the Progressive Matrices Test 9
BURT, C.	The Sign Pattern of Factor-Matrices : A Note on a Problem raised by Mr. Gabriel 15
HOWIE, D.	A Comparison of Two Methods of Factorizing Test Data 31
STUART, A.	The Correlation between Variate-Values and Ranks in Samples from a Continuous Distribution 37
WHITFIELD, J. W.	The Distribution of the Difference in Total Rank Value for Two Particular Objects in m Rankings of n Objects 45
NOTES AND CORRESPONDENCE	Limit Distributions for Total Rank Values 50
	The Identification of Factors from Different Experi- ments 52
	Spearman's Form of Criterion Analysis 57
	A Reliability Coefficient Based on Analysis of Variance 61
BOOK REVIEWS	 62

This *Journal* is issued by the British Psychological Society. The subscription price is 30s. a volume. Members of the Society should send their subscriptions to THE SECRETARY, British Psychological Society, Tavistock House, London, W.C.1 ; non-members to THE MANAGER, University of London Press Ltd., Warwick Square, London, E.C.4. Owing to the increased cost of paper and printing, the Society regrets that, until further notice, it will be possible to issue only two parts each year, namely, in May and November. The size of each volume will therefore be about 150 pages. Since the number of subscribers is increasing, it is hoped that in the near future it may be possible to increase the number to three parts annually.

Papers for publication should be sent to Sir CYRIL BURT, 9 Elsworthy Road, London, N.W.3. Contributors will receive 25 copies of their papers free. Extra copies may be purchased if the order is given when the proofs are returned. Contributors should indicate whether they wish their reprints to be bound in covers. The publishers will state the current prices when sending out proofs.

Authors are asked to send their papers in a form which is ready for submission to the printer : that is to say, the arrangement of the manuscript (including a brief list of contents at the head and a summary of results and conclusions at the end) should follow the general arrangement observed in previous publications of this *Journal*. A roneo'd pamphlet dealing with the preparation of papers in statistical psychology has been drawn up by the Editor and can be lent on request. Authors will be expected to pay the cost of corrections for which they are responsible.

Books for review should be sent to the Editor ; advertisements for insertion to the Manager, University of London Press, Ltd.

The British Psychological Society also issues quarterly *The British Journal of Psychology*, *General Section*, and *The British Journal of Medical Psychology*. Subscriptions, 40s. per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society also publishes, jointly with the Training College Association, *The British Journal of Educational Psychology*. Subscriptions, £1 per annum, for this Journal should be sent to Methuen & Co., Ltd., 36 Essex Street, London, W.C.2.

Papers for publication in the *General Section* should be sent to Professor D. W. HARDING, the Psychological Laboratory, Bedford College, London, N.W.1 ; those for publication in *The British Journal of Medical Psychology* should be sent to Dr. J. D. SUTHERLAND, 2 Beaumont Street, London, W.1 ; those for publication in *The British Journal of Educational Psychology* to Emeritus Professor C. W. VALENTINE, The White House, Wythall, Nr. Birmingham.

BRITISH JOURNAL OF EDUCATIONAL PSYCHOLOGY

Edited by C. W. VALENTINE

Vol. XXIV. Part. 1

J. J. B. DEMPSTER. Symposium on the Effects of Coaching and Practice in Intelligence Tests : III—Southampton Investigation and Procedure.

S. WISEMAN. Symposium on the Effects of Coaching and Practice in Intelligence Tests : IV—The Manchester Experiment.

E. A. PEEL. The Permanent Contribution of Francis Galton to Psychology.

E. L. BLACK. The Difficulties of Training College Students in Understanding what they Read.

K. P. WATTS. The Effects of a University Training in Mathematics on Scores in an Intelligence Test.

C. T. FORD. Developments in Written Composition during the Primary School Period.

SIR CYRIL BURT. Critical Notice—The Deprived and the Privileged (B. M. Spinley).

SUMMARY OF RESEARCH REPORTED IN DEGREE THESIS. A Study of Eight Primary Readers.

The Journal appears once each University term—in February, June and November.

Subscription : Twenty shillings per annum, post free, payable in advance, or 10s. per number.

Subscriptions should be sent direct to the Manager, B.J.E.P. Department

MESSRS. METHUEN & CO., LTD

36 Essex Street, Strand, London, W.C.2

Members of the British Psychological Society receive the *Journal* on special terms

HUMAN RELATIONS

A Journal of Studies towards the Integration of the Social Sciences

Contents of VOL. VI, No. 4

A. K. RICE. Productivity and Social Organisation in an Indian Weaving Shed.

R. B. CATTELL, DAVID R. SAUNDERS and GLEN F. STICE. The Dimensions of Syntality in Small Groups : 1. The Neonate Group.

J. M. M. HILL and E. L. TRIST. A Consideration of Industrial Accidents as a Means of Withdrawal from the Work Situation.

N. L. GAGE and RALPH V. EXLINE. Social Perception and Effectiveness in Discussion Groups.

MURRAY HORWITZ and DORWIN CARTWRIGHT. A Projective Method for the Diagnosis of Group Properties.

Book Reviews.

30s. per annum or 8s. 6d. per issue

The Tavistock Institute of Human Relations
London, England.

The Research Center for Group Dynamics
Ann Arbor, Mich.

Subscriptions and orders to Managing Editor

TAVISTOCK PUBLICATIONS LTD

2 Beaumont Street, London, W.1

THE BRITISH JOURNAL OF PSYCHOLOGY

Edited by D. W. HARDING

Vol. XLV. Part 1.

February, 1954

12s. 6d. net

- L. S. HEARNshaw. The psychological study of conceptual thinking.
HILARY M. CLAY. Changes of performance with age on similar tasks of varying complexity.
R. C. OLDFIELD. Memory mechanisms and the theory of schemata.
C. B. GIBBS. The continuous regulation of skilled response by kinaesthetic feed back.
D. SHEPPARD. The adequacy of everyday quantitative expressions as measurements of qualities.
E. C. POULTON. Verbal extrapolation and interpolation.
M. B. SHAPIRO. A preliminary investigation of the effects of continuous stimulation on the perception of 'apparent motion.'

CORRESPONDENCE

PUBLICATIONS RECENTLY RECEIVED.

The subscription price per volume, payable in advance, is 40s. net (post free)

Subscriptions may be sent to any bookseller or to the

CAMBRIDGE UNIVERSITY PRESS

BENTLEY HOUSE, EUSTON ROAD, LONDON, N.W. 1

CUSTODY OF UNPUBLISHED RESEARCH PAPERS

Arrangements have been made whereby the Technical Information and Documents Unit (T.I.D.U.) of the Department of Scientific and Industrial Research will, on the Editor's recommendation, accept custody of research papers (particularly those touching industry) for which, on account of their length, there is insufficient space in this *Journal*. This mainly concerns ancillary material which, being inessential to the general reader, need not occupy journal space, or possibly a complete paper of highly specialized interest. It is intended that specialist workers interested should have access to deposited material (either in the Reading Room of the Unit or by loan).

The Department of Scientific and Industrial Research (T.I.D.U.)

CUNARD BUILDING

15, Regent Street, London, S.W. 1

